Métodos quantitativos

Métodos quantitativos

Regis Garcia
Helenara Regina Sampaio
Regina Lúcia Sanches Malassise
Eliane Maria de Oliveira Araman
Marcelo Caldeira Viegas
André Marcelo Santos de Souza
Kiliano Gesser
Márcia Vilma Depiné Dalpiaz
Ana Luisa Fantini Schmitt
Débora Cristina Brandt

Diretor editorial e de conteúdo: Roger Trimer
Gerente de produção editorial: Kelly Tavares
Supervisora de produção editorial: Silvana Afonso
Coordenador de produção editorial: Sérgio Nascimento
Editor: Casa de Ideias
Editor assistente: Marcos Guimarães
Revisão: Mônica Santos
Capa: Bruno Portezan Jorge e Sheila Ueda Piacentini Barison
Diagramação: Casa de Ideias

Dados Internacionais de Catalogação na Publicação (CIP)

Malassise, Regina Lúcia Sanches
M238m Métodos quantitativos / Regina Lúcia Sanches Malassise, Marcelo Caldeira Viegas, Ana Luisa Fantini Schmitt, Débora Cristina Brandt, Regis Garcia, Márcia Vilma Depiné Dalpiaz, Kiliano Gesser, Helenara R. Sampaio, André Marcelo Santos de Souza, Eliane Maria de Oliveira Araman. – Londrina: Editora e Distribuidora Educacional S.A., 2014.
176 p.

ISBN 978-85-68075-12-8

1. Estatístico. 2. Probabilidade. I. Viegas, Marcelo Caldeira. II. Schmitt, Ana Luisa Fantini. III. Brandt, Débora Cristina. IV. Garcia, Regis. V. Dalpiaz, Márcia Vilma Depiné. VI. Gesser, Kiliano. VII. Sampaio, Helenara R. VIII. Souza, André Marcelo Santos de. IX. Araman, Eliane Maria de Oliveira. X. Título.

CDD 510

Sumário

Apresentação

Prezado acadêmico, a partir de hoje você inicia uma nova disciplina e com ela conhecerá o mundo dos métodos estatísticos e toda a sua relação com a estatística. Que tal iniciarmos com um breve histórico da estatística?

Para você ter ideia de como a estatística é antiga, estima-se que por volta de 5000 a.C. os registros de presos egípcios em documentos já continham dados estatísticos. Porém, é com a construção das pirâmides, já em 3000 a.C., que documentos comprovam na contagem o uso da estatística por causa da falta de mão de obra da época. Imagine você que naquela época já se fazia estimativa de quantidade de pessoas necessárias para terminar determinado trabalho.

Mas é na China que a estatística se coloca em posição de destaque na história, afinal, em 2238 a.C., quando o imperador ordenou um recenseamento para fins agrícolas e comerciais, era a estatística que permitia contabilizar os dados coletados. Com o passar do tempo os recenseamentos começaram a ficar bastante comuns, tanto que, em 600 a.C., os cidadãos do Egito tinham, todos os anos, que declarar sua profissão e sua fonte de rendimento.

Para você ter noção de como a estatística é importante na história, vou contar uma curiosidade sobre um recenseamento. Você conhece a Bíblia, não é mesmo? O livro conta, entre outras passagens, a história do cristianismo e é muito conhecido no nosso país. Pois, então, você sabia que a Bíblia dá relatos de estatística quando conta a história do nascimento de Jesus? Naquela época, por ordem do senado, todos os indivíduos deveriam retornar ao seu local de nascimento para um recenseamento, e foi nessa viagem que nasceu Jesus, justamente quando José e Maria, seus pais, estavam voltando à terra natal de José para participar do censo.

Sendo assim, podemos dizer com toda a certeza que a estatística nasceu com o objetivo de coletar, contar e organizar dados, sempre procurando pesquisar todos os indivíduos. Mas é claro que a palavra <u>estatística</u> não existia

em épocas tão distantes, isto é algo mais recente! Ela foi usada pela primeira vez pelo alemão Gottfried Achenwall, em 1752, e deriva da palavra latina *status,* que significa estado, numa alusão ao aproveitamento que os governantes estavam fazendo da estatística. Tudo a ver com o recenseamento, não é mesmo?

Daí você deve estar se perguntando, mas onde entram os tais métodos estatísticos? Os métodos estatísticos fazem parte da estatística, pois servem para fazer o tratamento dos dados coletados nas mais diversas situações. Por exemplo, vamos supor que uma empresa está se preparando para lançar um novo produto e precisa conhecer as preferências dos consumidores no mercado de interesse em que pretende atuar. Para isso, esta empresa pode fazer uma pesquisa de mercado entrevistando um número de residências escolhidas aleatoriamente. E poderá usar os resultados para estimar as preferências de toda a população. Para isso é provável que esta empresa tenha coletado os dados e utilizado algum método quantitativo, como o desvio-padrão, o coeficiente de variação ou mesmo a regressão linear para saber ou estimar a quantidade de futuros compradores do produto que quer lançar.

Enfim, tudo isso você vai estudar ao longo da disciplina, de maneira gradual, passando pelos conceitos básicos, pelas medidas de posição e de dispersão, terminando por fazer regressões com o intuito de comparar dados. Claro que neste ínterim você também vai aprender sobre os gráficos e as distribuições de frequência, que permitem a visualização dos dados coletados.

Gostou? Então pode esperar muito mais! A disciplina lhe dará muitos subsídios para trabalhos de pesquisa, futuros artigos e projetos!

Bons estudos!

Unidade 1

Princípios da estatística

Regina Lúcia Sanches Malassise
Eliane Maria de Oliveira Araman
Helenara Regina Sampaio
Regis Garcia

Objetivos de aprendizagem: Nesta unidade, você será levado a compreender os principais conceitos da estatística bem como sua relação com o meio corporativo e, ao final da leitura, terá condições de reproduzir as aplicações básicas da estatística em seu ambiente de trabalho.

Seção 3: **Organização e deflação de dados**

Nesta seção, apresentaremos as principais formas de organização e complementaremos com estudos sobre deflação de dados.

Introdução ao estudo

Nos dias de hoje, muitas informações que nos chegam sobre o comportamento das pessoas em relação aos mais variados temas são frutos de pesquisas que, com certeza, receberam algum tipo de tratamento estatístico. Às vezes, até comerciais procuram estimular nossa curiosidade por meio de algum questionamento estatístico. Talvez você já tenha visto a propaganda que fala: "A ciência ainda não sabe explicar por que um ser humano boceja quando outro boceja ao seu lado." Então, a pergunta, em si, é uma afirmação – se você boceja, seu colega ao lado vai bocejar também –, mas ela só foi possível de ser feita porque a observação desse fato foi acompanhada, tabulada, e demonstrou que, na grande maioria dos casos observados, bocejar era comum entre pessoas que estavam próximas umas das outras.

Bom, mas o objetivo aqui não era levá-lo a ficar com sono... e, por favor, mantenha os olhos bem abertos para poder aprender cada vez mais. Um questionamento interessante para você seria: "Por que as pessoas resolvem estudar Administração de Empresas?" Com certeza, as pessoas apresentarão os mais diversos motivos, então tais motivos devem ser agrupados por itens em comum e tabulados para, depois, fornecerem informações – por exemplo, em percentuais — que possam revelar os motivos mais importantes que levaram o grupo de pessoas pesquisadas a estudar Administração. E, assim, muitas questões instigantes necessitam da aplicação de conhecimentos estatísticos para serem respondidas. Nesse sentido, o conteúdo desta unidade está dividido em três seções: Seção 1 — Introdução à Estatística e Fases de sua Aplicação, Seção 2 — A Informatização da Estatística e Seção 3 — Organização e Deflação de Dados. Logo, nesta unidade vamos tratar dos conceitos básicos necessários à compreensão da estatística como área do conhecimento humano e, sobretudo, como fornecedora de ferramentais práticos que possibilitam a solução de problemas cotidianos dos mais variados. O foco está na revisão dos principais conceito se no entendimento da relação entre a disciplina e a realidade das empresas.

Seção 1 Introdução à estatística e fases de sua aplicação

Caro aluno, creio que você, toda vez que inicia o estudo de uma nova disciplina, tenha necessidade de conhecê-la melhor, eessa necessidade aumenta quando tal disciplina envolve uma área do conhecimento diferente daquela que sustenta sua formação principal. Pois bem, esse é o caso da disciplina de Estatística, que é a base estrutural de Métodos Quantitativos. Assim, nesta seção vamos juntos conhecer um pouco mais sobre a história da Estatística, seus conceitos básicos e as etapas de aplicação. Vamos lá! Vamos dar início aos nossos estudos.

1.1 Evolução da estatística

A estatística tem suas primeiras aplicações associadas às necessidades do Estado de ter uma fonte de informações agrupadas para poder formular políticas públicas. Assim, dados econômicos e demográficos davam suporte aos estudos da administração pública — em especial para fins militares e tributários. Os momentos mais importantes são:

- A obra pioneira de Francesco Sansovini, de 1561, sobre estatística descritiva e o reconhecimento por parte da Igreja Católica Romana da importância de manter-se arquivos com registros de batismos, casamentos e óbitos — e que se tornaram obrigatórios a partir do Concílio de Trento (1545-1563);
- Posteriormente, os estudos de Gottfried Achenwall (1719-1772), professor da Universidade de Göttingen e a quem se atribui a criação do termo “estatística”, aprimoraram a sistematização e definição da mesma orientação descritiva dos estatísticos italianos;
- A estatística começa a encaminhar-se para aquilo que seria seu futuro como “ciência do significado e uso dos dados” (MEMÓRIA, 2004, p. 12). Como estudo independente, começa a ganhar espaço com o livro de John Graunt de 1662.

> Os dados usados por Graunt compreendiam uma série anual de 1604 a 1660, coletados nas paróquias de Londres, de onde ele tirou as seguintes conclusões: que havia maior nascimento de crianças do sexo masculino, mas havia distribuição aproximadamente igual de ambos os

> sexos na população geral; alta mortalidade nos primeiros anos de vida; maior mortalidade nas zonas urbanas em relação às zonas rurais (MEMÓRIA, 2014, p. 13).

- Em 1683, William Petty denominou a estatística como Aritmética Política, que para ele era a arte de raciocinar observando os dados sobre os fatos relacionados ao governo;
- Edmond Halley construiu a primeira tabela de sobrevivência, que serviu de base para o cálculo atuarial — em especial para o seguro de vida;
- Blaise Pascal (1623-1662) e Pierre de Fermat (1601-1665) investigaram a questão das probabilidades visando solucionar problemas relacionados aos jogos de azar, de tal forma que:

> Os primeiros problemas sobre probabilidades refletiram o desenvolvimento da análise combinatória em jogos de azar. Em todos eles eram examinados os diferentes modos em que arranjos e combinações podiam ser empregados na enumeração dos casos favoráveis. Esses problemas eram dominados por considerações sobre os casos igualmente prováveis, com as probabilidades determinadas *a priori*, onde foi utilizado o seguinte tipo de raciocínio: dado uma urna contendo a (bolas pretas) e b (bolas brancas), a probabilidade de se extrair uma bola preta é igual aa/(a+b) (MEMÓRIA, 2014, p. 15).

- Em especial, os feitos registrados ao longo dessa evolução podem ser resumidos nos momentos mais importantes, conforme destacou Castro (1967 apud GARCIA, 2010, p. 2-3):

> O primeiro período se caracteriza pelos registros sistemáticos e cadastros das informações do Estado com finalidades principais ligadas à guerra ou a interesses fiscais. Esse modelo vai desde o período feudal até meados do século XVII. O segundo período dá à estatística o caráter de disciplina autônoma e se referia às análises de registros de batismos, casamentos e enterros. Ainda segundo Castro (1967), em 1949 a estatística foi definida em termos de objeto e relações com as ciências por Gottfried Achenwall. O terceiro período constitui a evolução da estatística até os dias de hoje, caracterizada por constantes aperfeiçoamentos técnicos e científicos.

Pois bem, do exposto até aqui, chega o momento em que podemos pontuar qual é o objeto e o objetivo da Estatística. Podemos dizer que o **objeto da**

estatística é o estudo de fenômenos que possam ser quantificados de alguma forma por meio da observação. O **objetivo da estatística**, por sua vez, é analisar os dados obtidos em uma pesquisa e, por isso mesmo, tecer ligações de causalidade existentes entre eles — mas essas ligações só se estabelecem com sentido correto quando se utilizam os conhecimentos científicos das áreas às quais essa pesquisa se refere.

Dessa forma, a estatística encontra aplicações em quase todos os campos da atividade humana. O administrador necessita conhecer o mercado de um produto ou serviço e o faz analisando dados de pesquisas de mercado; o gestor público precisa compreender as finanças públicas e, muitas vezes, utiliza-se de séries históricas que recebem algum tratamento estatístico como forma de resumir e expressar o que de fato o grupamento de dados quer dizer; os agrônomos e meteorologistas necessitam conhecer as condições climáticas para entender seus reflexos sobre o tempo e a agricultura; os biólogos e os profissionais envolvidos nos segmentos de saúde precisam estudar as epidemias e, a partir de projeções estatísticas, podem prevê-las e antecipar-se aos seus efeitos danosos na população. Enfim, além dessas, uma série de outras aplicações dos conhecimentos estatísticos são possíveis, de tal forma que:

> Atualmente, conforme comentado na seção de introdução, os meios informatizados dos quais dispõem os indivíduos e as organizações possibilitam que aplicações de alta complexidade sejam desenvolvidas a partir de um conjunto mínimo de conhecimentos sobre o tratamento de dados. Dentre os conjuntos de técnicas capazes de oferecer ferramentais eficazes para esses tratamentos, destacam-se os estatísticos, geralmente muito bem fundamentados e testados empiricamente. Esses métodos são desenvolvidos e testados no ambiente acadêmico e de pesquisa de diversas áreas, tais como economia (por meio da econometria), contabilidade (por meio da contabilometria) e, principalmente, nas áreas das ciências biológicas, em especial pela utilização das técnicas que permitem o estudo de causa e efeito das doenças (GARCIA, 2010, p. 3-4).

Dessa forma, apresentamos no Quadro 1.1 a interligação que existe entre a estatística e outras áreas do conhecimento, de tal forma que o uso de métodos estatísticos permite compreender melhor os problemas das mais diversas áreas.

Quadro 1.1 Métodos estatísticos aplicados em diversas áreas

Fonte: Garcia (2010, p. 5).

Questões para reflexão

Você sabia que existe um levantamento sobre a taxa de sobrevivência das pequenas empresas no Brasil? Se sim, você sabe dizer qual é sua atual estatística?

Para que a estatística contribua para o desenvolvimento de pesquisas, ela necessita lançar mão de um conjunto de métodos que permitam coletar, agrupar, resumir e analisar dados para, então, apresentar os resultados obtidos pela análise das variáveis observadas. É, pois, a partir da etapa de análise que é possível obter conclusões válidas para o processo de tomada de decisão. Porém, antes de falarmos de método estatístico, é necessário que nos familiarizemos com o entendimento de alguns conceitos básicos da estatística.

1.1.1 Estatística: termos usuais

Partindo do entendimento de que a estatística constitui-se em um conjunto de técnicas que objetivam descrever, analisar e interpretar os dados numéri-

cos de uma população ou amostra, podemos, então, destacar alguns termos importantes e comumente utilizados nessa área. São eles:

a) **População**: refere-se ao conjunto de elementos que apresentam uma característica em comum e com o qual se deseja fazer um estudo — também chamado de universo de dados. Nas pesquisas, deve-se definir claramente a população para que se possa delimitar o que se pretende conhecer. Por exemplo, na nossa questão para reflexão, comentamos sobre taxa de sobrevivência de pequenas empresas no Brasil. Bom, nossa população, na pesquisa, será todas as pequenas empresas brasileiras. Por outro lado, o primeiro passo será pesquisar na literatura para saber qual é o critério de classificação de pequenas empresas (qual tipo), pois, assim, quando encontrarmos essa taxa, teremos certeza do tipo de pequena empresa a que ela se refere. Comentamos isso para que você fique atento a esse detalhe, pois, no Brasil, temos, no mínimo, dois critérios de classificação de pequenas empresas: um de acordo com o número de funcionários (critério SEBRAE) e outro de acordo com o faturamento mínimo anual (critério do Banco do Brasil). Observe, ainda, que cada pesquisa em particular pode tomar um critério de população já existente ou definir sua população de acordo com os interesses de sua pesquisa. Um exemplo de pesquisa que envolve toda a população é o Censo Demográfico, que é realizado decenalmente pelo Instituto Brasileiro de Geografia e Estatística.

Para saber mais

Você pode conhecer mais sobre os dados do Censo 2010 consultado a página do IBGE no link abaixo:

<http://www.ibge.gov.br/home/estatistica/populacao/censo2010/>.

b) **Unidade:** refere-se a qualquer elemento ou indivíduo da população. Por exemplo, em relação aos resultados do Censo, uma entrevista com Ana Nery da Silva apresenta as informações de escolaridade dessa mineira de 38 anos que concluiu o segundo grau completo. Essa mulher é uma unidade que foi entrevistada pelo Censo.

c) **Amostra**: é um subconjunto da população, ou seja, é uma parte da população. É sempre mais utilizada nas pesquisas do que a população. Isso ocorre porque a população envolve um número muito grande de observações; logo, a amostra é uma forma de trabalhar de maneira confiável, com valores que expressam o que ocorre com a população. Por exemplo, em relação ao Censo, poderíamos estar interessados em identificar o grau de escolaridade dos jovens de 28 anos do município de Belo Horizonte. Estaríamos, então, destacando um grupo de jovens específicos e extraindo os dados de uma população do Censo. Esse grupo de jovens seria, então, uma amostra.

d) **Variável:** refere-se à informação ou ao atributo que é o objeto de estudo. Por exemplo, numa pesquisa sobre os universitários, posso pesquisar o sexo. As variáveis podem ser **qualitativas**, que qualificam os dados — por exemplo, estado civil, sexo e escolaridade —, e podem ser **quantitativas**, isto é, podem ser mensuradas — como, por exemplo, peso, altura etc. Na variável quantitativa, os dados podem ser expressos por meio de uma estrutura numérica, isto é, que podem ser enumerados — quantidade de objetos, idades, tamanho etc. A variável quantitativa, ainda, pode ser contínua ou discreta.

e) **Variável discreta** ou descontínua: são os valores que obtemos por meio das contagens, como, por exemplo, o número de alunos de uma sala de aula do 1º ano do Ensino Fundamental em 2010 = 30, em 2011 = 33, em 2012 = 32 e em 2013 = 35. Seus valores são expressos, geralmente, por meio de números inteiros e não negativos.

f) **Variável contínua**: os valores obtidos por meio dessa mensuração correspondem ao conjunto dos números reais, de forma que podem assumir, teoricamente, qualquer valor entre dois limites. Exemplos: as horas de um curso, temperatura etc.

O Quadro 1.2 apresenta um resumo sobre a classificação das variáveis.

Quadro 1.2 Classificação das variáveis

Variáveis	Tipos	Descrição	Exemplos
Qualitativas ou Categóricas	Nominal	Não existe nenhum tipo de ordenação dos dados.	Sexo, cor dos olhos, cor da pele, estado civil.
	Ordinal	Obedece a um certo tipo de ordenação.	Classe social, grau de instrução, etc.
Quantitativas	Discretas	Dados são obtidos por meio de contagem.	Número de funcionários de uma empresa; número acidentes de trânsito ocorrido durante um mês.
	Contínuas	Dados são obtidos por meio de medição.	Medidas de altura e peso das pessoas.

Fonte: Araman e Sampaio (2009, p. 4).

g) **Dados brutos:** são os dados que se apresentam da forma como foram coletados e que ainda não foram organizados.

h) **Análise dos dados**: após a coleta dos dados, é necessário realizar a análise dos mesmos. Os resultados da análise podem ser expostos em tabelas de forma sintética e submetidos, ou não — de acordo com cada pesquisa —, a um tratamento estatístico mais profundo, em que todas as informações reunidas nos passos anteriores são comparadas entre si e analisadas.

i) **Tabelas estatísticas**: são dados agrupados em colunas e linhas. Existem regulamentações para a construção de tabelas estatísticas, mas, em linhas gerais, ela deve seguir algumas regras básicas: deve ser clara e objetiva e, ainda, deve demonstrar o comportamento das variáveis através de representações simples que possibilitem ao leitor ter uma perfeita compreensão do fenômeno sem muito esforço. As tabelas estatísticas podem ser apresentadas em série cronológica ou temporal, série de localização ou geográfica, série específica e série de distribuição de frequência.

j) **Gráficos**: constituem a representação gráfica das séries estatísticas de uma pesquisa. Têm por finalidade representar os resultados obtidos, permitindo apontar conclusões acerca da evolução do fenômeno observado ou da forma como este se relaciona com os valores da série. Existem diferentes tipos de dados, e hoje os softwares como o Excel apresentam diferentes formas de composição dos mesmos. O essencial, para o pesquisador, é ter em mente que os gráficos devem prezar pela simplicidade, clareza e veracidade dos fatos que representam.

k) **Análise exploratória:** o termo exploratório é sugestivo, tendo analogia com a palavra "desbravar", ou seja, pesquisar algo desconhecido, explicitando novas descobertas. Ela emprega grande variedade de técnicas gráficas e quantitativas (inclusive a análise de regressão) e objetiva maximizar a obtenção de informações ocultas na estrutura dos dados e a descoberta de variáveis importantes. Observa-se que, nesse caso, há uma profundidade maior em relação à análise dos dados, ou seja, não há uma preocupação meramente voltada para o conhecimento sobre as características principais dos dados, mas, sim, para a busca por informações ocultas que resultem em explicações, por exemplo, sobre causas e efeitos entre as variáveis envolvidas. Com os softwares disponíveis, essa técnica possibilita a descoberta de quais tendências, relações e padrões poderiam estar ocultos em uma coleção de dados analisados. A aplicação de técnicas adequadas implica a análise minuciosa do problema e dos resultados pretendidos.

l) **Análise descritiva:** menos pretensiosa do que a exploratória, a análise descritiva se volta para o conhecimento dos dados em relação às suas principais características. Isso está longe de significar menor importância para esse tipo de análise. Ao contrário, a diferença de profundidade e extensão visa apenas adequar o tratamento às reais necessidades do analista. É importante que haja sempre o questionamento sobre a adequação, ou não, dos procedimentos utilizados em termos de resultados para o usuário. Não é eficaz um tratamento que gere mais informações do que o necessário, e tampouco aquele que as gere com deficiência. A análise descritiva implica o cálculo de parâmetros que possibilitem a compreensão das características principais de um grupo de elementos. Entre essas características, podem-se destacar a soma, a média, a variância, o desvio-padrão, a mediana, a moda etc.

Para saber mais

Você deve estudar mais sobre estatística descritiva lendo o Capítulo 2 — "Estatística Descritiva" — do livro **Estatística aplicada**, dos autores Ron Larson e Betsy Farber, publicado em 2004.

Conhecidos os principais conceitos ligados à estatística, vamos saber um pouco mais sobre amostragem e método estatístico.

1.1.2 Amostragem com método probabilístico

A teoria da amostragem garante que se possa fazer a inferência, ou seja, a extrapolaçãodo que se observou na amostra para a população da qual ela foi retirada. Do contrário, haveria grande possibilidade desse incorrer em erro. Um aspecto importante para uma pesquisa científica é a definição da amostragem. Existem algumas formas de se fazer isso, e vamos ver a seguir algumas delas.

Métodos probabilísticos: nesse método, cada elemento da população deve possuir uma mesma probabilidade de ser selecionado. Então, se N corresponde ao tamanho da população, 1/N corresponde à probabilidade de cada elemento ser selecionado. Esse método garante cientificamente a aplicação da inferência estatística, ou seja, somente com base nas amostras feitas por métodos probabilísticos é que podemos fazer inferências a respeito da população a partir da análise da amostra. Além disso, essa técnica garante o acaso na escolha.

Amostragem casual ou **aleatória simples:** esse processo de amostragem é o mais simples e também o mais utilizado. Para formar a amostra, numeramos a população e realizamos um sorteio aleatório por meio do qual qualquer elemento da população possa ser sorteado. Por exemplo, se quisermos compor uma amostra de 10% dos 300 professores desta universidade para uma determinada pesquisa, devemos, primeiramente, numerar os professores de 1 a 300. Depois disso, colocamos todos os números em uma urna, misturamos bem, e retiramos, um a um, trinta números que irão compor a amostra. Esse tipo de amostragem deve ser usado sempre que se tratar de uma população homogênea, ou seja, quando há uma distribuição uniforme dos dados na população. O número de elementos para compor uma amostra dependerá de vários fatores, como tempo para a pesquisa, verba etc. Uma das formas de fazer esse tipo de amostragem é tomar cerca de 5% a 10% dos elementos da população, pois quanto maior a amostra, maior a precisão.

Amostragem proporcional estratificada: esse tipo de amostragem é adequado quando a população é dividida em estratos ou subgrupos de população. Nesse caso, convém que o sorteio dos elementos que irão compor a amostra considere os estratos da população, de forma que tenhamos uma amostra significativa em cada estrato. Consideremos o exemplo anterior dos professores da universidade. Vamos obter uma amostra estratificada de 10%, supondo que, entre os 300 professores, 170 sejam do sexo masculino e 130, do sexo feminino. Dessa forma, temos dois estratos, um para o sexo masculino e outro para o sexo feminino. Veja como ficaria essa amostra no Quadro 1.3. Quando realizamos

uma amostragem estratificada, garantimos a representatividade de cada estrato da população, o que já não podemos mais garantir pela amostragem casual.

Quadro 1.3 **Amostra estratificada dos professores da universidade**

Sexo	População	10%	Amostra
Masc.	170	17	17
Femin.	130	13	13
Total	300	30	30

Fonte: Araman e Sampaio (2009).

Amostragem sistemática: utilizamos esse tipo de amostragem quando os elementos da população já se encontram ordenados ou organizados de alguma maneira. Dessa forma, não é necessário construir o sistema de referência. Analise o seguinte exemplo: vamos supor que um conjunto residencial seja formado por 216 apartamentos e que desejamos realizar uma pesquisa de opinião lá. Para compor uma amostra composta por 36 apartamentos, podemos usar o seguinte procedimento:

1. Se desejamos formar uma amostra com 36 apartamentos em um total de 216, então temos 216/36 = 6.
2. Em seguida, escolhemos por sorteio um número entre 01 e 06 e que será o primeiro elemento da amostra.
3. Os demais elementos seriam considerados de seis em seis a partir desse primeiro.
4. Assim, suponhamos que o número sorteado para ser o primeiro tenha sido o 4. Então, a amostra seria composta pelo 4º apartamento, 10º apartamento, 16º apartamento, 22º apartamento, e assim sucessivamente em uma ordem estabelecida. **Amostragem por julgamento:** esse tipo de amostragem acontece quando o pesquisador utiliza seu próprio julgamento para selecionar os membros de uma população que oferecem boas perspectivas de fornecimento de informações necessárias para a sua pesquisa.

Amostragem por quotas: a amostragem por quotas é realizada quando o pesquisador encontra e entrevista um número predeterminado de pessoas pertencentes a cada uma das várias categorias da população.

1.1.3 Métodos não probabilísticos

Nos métodos não probabilísticos, os elementos que irão compor a amostra são escolhidos de forma deliberada. Nesses casos, não é possível fazer a generalização dos resultados obtidos na pesquisa, uma vez que as amostras não probabilísticas não garantem a representatividade da população. Vejamos alguns casos.

Amostragem acidental: nesse tipo de amostragem, a amostra é formada por aqueles elementos que vão aparecendo, ou seja, que são possíveis de se obter até completar-se o número certo de elementos daquela amostra. Temos, como exemplo, as pesquisas de opinião realizadas em praças públicas, ruas do centro das cidades etc. Nesses casos, os entrevistados são acidentalmente escolhidos para compor a amostra.

Amostragem intencional: esse tipo de amostra é definido pelo pesquisador após ter-se estabelecido um determinado critério. A partir desse critério, o pesquisador escolhe intencionalmente um grupo de elementos para compor a amostra. Então, o pesquisador, de forma intencional, dirige-se diretamente a grupos de elementos dos quais deseja coletar seus dados. Por exemplo, ao realizar uma pesquisa a respeito da preferência por um determinado cosmético, o pesquisador vai intencionalmente a um (ou vários) salão de beleza e entrevista as pessoas que ali se encontram naquele momento.

Amostragem por quotas: esse método de amostragem é muito utilizado em pesquisas de mercado e também em prévias eleitorais. Para se fazer uma amostragem por quotas, primeiramente, é necessário realizar a classificação da população em relação a algumas propriedades que podem ser consideradas relevantes para os objetivos da pesquisa. Depois disso, é necessário determinar a proporção da população de acordo com cada característica, baseada na constituição estimada, presumida ou mesmo já conhecida da população. Em seguida, fixamos cotas para cada entrevistador que ficará responsável por selecionar os entrevistados, de forma que a amostra total dos entrevistados respeite a proporção já determinada anteriormente. Por exemplo, em uma pesquisa, desejamos conhecer mais a respeito do trabalho das mulheres no mundo moderno. Nessa pesquisa, provavelmente, será interessante considerar vários aspectos, como o trabalho na cidade e no campo, o número de filhos, a idade em que os filhos se encontram etc. Estabelecemos, então, quotas para que cada entrevistador selecione e entreviste o número de mulheres estabelecido em cada classe que deve ser estudada.

Atividades de aprendizagem

1. Observe as situações descritas e classifique-as de acordo com o método de amostra não probabilístico utilizado.
   a) Um estudante cuja curiosidade era verificar os números de acessos as redes sociais de um grupo de jovens, escolhe entrevistar seus colegas de sala de aula.
   b) Para realizar uma pesquisa sobre o interesse das mulheres em realizar um curso superior no município X, uma empresa verificou que, no município em questão, existiam 13.800 mulheres entre 18 e 45 anos. Ela determinou, então, que cada entrevistador realizaria a pesquisa com 5 mulheres de 18 anos, 4 mulheres de 22 anos, 3 mulheres de 25 anos, 2 mulheres de 30 anos e 1 mulher de 40 anos.
2. Diferencie a amostragem determinada por métodos probabilísticos e por métodos não probabilísticos.

1.1.4 Cuidados com a amostragem

Uma boa pesquisa depende bastante da amostra selecionada, que deve garantir a representatividade da população. Diante disso, para que não haja erros na amostragem, é importante que o pesquisador observe:

Definição do universo: a população, ou universo, deve ser definida de acordo com os objetivos da pesquisa etambém em concordância com o problema que se deseja investigar.

Definição da unidade da amostra: o pesquisador deve definir, também, a amostra que será a base do processo da seleção e investigação. Por exemplo, se deseja que sua amostra seja composta por famílias, então deve definir bem o que vai considerar como família e somente entrevistar aqueles que realmente se ajustam na definição de família adotada na pesquisa, de forma que outros grupos que não correspondam à definição pré-estabelecida não façam parte da amostra.

Confiabilidade: o rigor metodológico e os cuidados procedimentais é que vão garantir, ou não, a confiabilidade da pesquisa.

Tamanho da amostra: existem várias fórmulas para se calcular o tamanho adequado de uma amostra. Entretanto, devemos sempre levar em conta o tamanho de uma população. Normalmente, quanto maior for o tamanho da amostra (quanto maior for o número de elementos que compõem a amostra), menor serão os desvios dos parâmetros em relação ao que se esperava da população. Outro aspecto importante para consideramos é a homogeneidade da população — quanto mais homogênea for a população, menor será a amostra a ser pesquisada.

Determinando o tamanho da amostra

Há três fatores que determinam o tamanho da amostra:

- O grau de confiança adotado.
- O máximo erro permissível.
- A variabilidade da população.

Dessas considerações é que deriva a fórmula para o cálculo do tamanho da amostra, nesse momento considerando populações infinitas.

$$n = \left(\frac{z.\sigma}{E}\right)^2$$

Na fórmula anterior, podemos identificar os elementos que determinarão o tamanho necessário para que a amostra atenda ao nível de confiança desejado para as inferências sobre a população.

Para determinarmos, então, o tamanho da amostra para populações infinitas, primeiro encontramos o valor de z (grau de confiança adotado) com base na tabela de distribuição normal. Necessitamos do desvio-padrão populacional (variabilidade da população), que pode ser obtido por estimativa quando não conhecido, e finalmente do erro tolerável (o máximo erro permissível), determinado para o caso em estudo de acordo com a necessidade de precisão apresentada.

Para exemplificar a aplicação dessa fórmula, imaginemos uma empresa que deseja analisar a média de consumo de alimentos em seu refeitório no mês de dezembro. Em outros estudos feitos anteriormente, o desvio-padrão de consumo foi de 80 gramas. Deseja-se estimar o padrão de consumo com 95% de confiança e com erro admissível de 20 gramas. Quantas refeições deverão ser medidas e anotadas para a análise? Substituindo os valores da fórmula (7) tem-se:

$n = \left(\frac{1,96 \ x \ 80}{20}\right)^2$ = 61,465– o que equivale a aproximadamente 62.

No exemplo apresentado, seriam necessárias 62 refeições na amostra para se determinar o consumo médio na refeição durante o mês de dezembro. Essa forma de se calcular o tamanho da amostra é aplicada, conforme mencionado, nos casos em que a população é infinita, ou seja, não se tem um limite superior definido.

1.2 Fases da aplicação estatística

Como em todo método, existem fases devidamente identificáveis que permitem que se obtenha sucesso nos objetivos traçados. A estatística, conforme Fernandes (1999), tem como objetivo fornecer informação (conhecimento) utilizando quantidades numéricas. Seguindo esse raciocínio, de forma resumida, pode-se dizer que ela divide o estudo e a análise dos dados (fatos numéricos) em três fases:

1. Obtenção dos dados.
2. Descrição, classificação e apresentação dos dados.
3. Conclusões a tirar dos dados.

No caso das empresas, foco principal deste capítulo, a obtenção dos dados se dá de várias formas. Muitos são frutos do sistema de informação organizado pela empresa de acordo com seus objetivos informacionais. Outros são coletados via pesquisas de mercado, internamente ou de forma terceirizada, de acordo com a finalidade da análise que se deseja. Quando os dados são tirados do próprio sistema informacional da empresa, certamente há uma economia de recursos. Historicamente, o que se faz de forma planejada e antecipadamente pensada evita os gastos extraordinários gerados toda vez que se necessita de uma análise de algo sobre o qual não se tinha um objetivo informacional prévio.

Outra questão importante no processo de obtenção de dados é saber qual a quantidade suficiente para cobrir toda a necessidade da análise que se procederá. Aproveitar o processo de coleta para atender à real demanda é imprescindível para a economia de recursos. Por isso, seja interna ou terceirizada, seja planejada ou contingencial, a coleta de dados deve ser muito bem pensada e discutida entre os envolvidos no processo.

Uma vez obtidos os dados, estes passam por um processo denominado descrição, classificação e apresentação. A descrição e a classificação referem-se à

organização e ao tratamento dos dados, de forma a evidenciar inteligivelmente cada tipo de dado e aquilo que é possível ser útil para as conclusões às quais se pretende chegar. É nessa fase que se encontram os parâmetros estatísticos que serão estudados sob o enfoque da estatística descritiva.

Após a descrição dos dados, outras técnicas podem ser empregadas a fim de possibilitar um aprofundamento na análise e a descoberta de relações entre variáveis ou entre fenômenos.

Exemplo de análise descritiva de dados

Todos os anos, o INEP realiza o Censo Escolar, que é disponibilizado na forma de microdados agrupados por escola; docentes, turma e matriculados. Esses dados têm por objetivo fornecer informações estatísticas e servir de subsídio para a adoção de políticas e estratégias educacionais. Em 2007, realizou-se um levantamento tomando por base os Microdados do Censo Escolar, o qual trouxe o número de alunos matriculados bem como o de aprovados na modalidade Ensino de Jovens e Adultos (EJA) de 1ª à 4ª série registrados no agrupamento da escola. Apresentaremos abaixo algumas características da infraestrutura das escolas que ofereciam ensino nessa modalidade. A análise de dados abaixo tem como objetivo identificar variáveis da infraestrutura física e administrativa das escolas que oferecem EJA.

1.2.1 Descrição das variáveis selecionadas sobre a escola

De acordo com os dados, em 2007 o Brasil possuía 81.635 escolas que ofereciam a modalidade de ensino EJA (População).

Começamos nossa análise trabalhando com a infraestrutura escolar somente das escolas que têm alunos na quarta série (Amostra). Vamos fazer uma descrição da infraestrutura física (pública, interna e equipamentos) que essas escolas têm.

Os dados brutos informam que as escolas que possuem alunos na quarta série são 3.676. Dessas escolas, 828 localizam-se na zona rural e 2.848 na zona urbana. Juntas, ela apresentam um total de 36.533 salas de aula e contam com 68.376 funcionários. Naquele ano, registraram um total de alunos aprovados/concluintes da quarta série de 62.715.

Analisando as informações sobre a infraestrutura das escolas (conforme o Quadro 1.4), observa-se que a grande maioria das escolas conta com uma boa estrutura fornecida pelos serviços públicos. Os destaques são para a energia elétrica fornecida para 98% das escolas, a coleta de lixo em 85% delas, o fornecimento de

água tratada da rede pública que atende 82% das escolas e, por último, destacam-se os serviços de esgoto, que atendem apenas 56% das escolas pesquisadas.

Quadro 1.4 **Infraestrutura de serviços públicos básicos de que dispõem as escolas com a modalidade de ensino EJA de 1ª à 4ª série**

Tipo	Escolas que têm	% do total de escolas
Água de rede pública	3026	82,32%
Energia elétrica	3612	98,26%
Provida de esgoto	2088	56,80%
Coleta de lixo regular	3042	85,75%

Fonte: Dos autores (2014).

O Quadro 1.5 destaca a infraestrutura interna de cada escola, isto é, o que é oferecido na escola. Destaca-se que 96% das escolas funcionam em uma instalação classificada como prédio escolar, ou seja, uma construção destinada a abrigar salas de aula. Verificamos que apenas 65,5% das escolas contam com sanitários no próprio prédio.

Quadro 1.5 **Infraestrutura interna das escolas com modalidade de ensino EJA de 1ª à 4ª série**

Tipo	Escolas que têm	% do total de escolas
Prédio escolar	3555	96,71%
Sanitário no prédio da escola	2406	65,45%
Antena parabólica	1384	37,65%
Quadra de esportes	1605	43,66%
Biblioteca	1945	52,91%
Acesso à internet	1750	47,61%
Oferece alimentação	3538	96,25%
Oferece 5ª série EJA	1696	46,14%

Fonte: Dos autores (2014).

Outro destaque é que a maior parte das escolas oferece alimentação — 96% delas —; há biblioteca em 52% das escolas, acesso à internet em 47% e quadra de esportes em 43%. E em 46% das escolas oferece-se a 5ª série.

Quadro 1.6 **Equipamentos e recursos das escolas com modalidade de ensino EJA de 1ªà 4ª série**

Tipo	Escolas que têm	% do total de escolas
Aparelho de televisão	3088	84,00%
Aparelho de vídeo	2614	71,11%
Computador para aluno	1428	38,85%

Fonte: Dos autores (2014).

O Quadro 1.6 destaca a existência de equipamentos na escola. A maior parte das escolas possui televisão (84% delas) e aparelho de vídeo (84%). Porém, os computadores disponíveis para os alunos são oferecidosem apenas 38,8% das escolas.

1.2.2 Benefícios do uso da estatística

Certamente, ao pensar em uma análise complexa — por exemplo, em um controle de qualidade de produtos que são fabricados aos milhões em uma determinada organização, que possuem muitos itens a serem testados e que, em muitos casos, necessitam de testes que implicam a destruição do elemento para a realização da análise —, surgem questionamentos importantes do tipo: "Qual a melhor forma de se garantir que haja segurança nas afirmações sobre a qualidade de um lote de produtos nessas condições?"; "Qual a melhor forma de se utilizar o mínimo de recurso para se obter um resultado eficaz?"; "Qual a melhor forma de fazer com que os resultados esperados sejam obtidos no menor tempo possível?".

Esses questionamentos são feitos a todo o momento no contexto corporativo, principalmente diante da situação de acirramento da competição.Para Stock e Watson (2004, p. 39), a estatística "utiliza dados para entender o mundo que nos cerca". Entender o comportamento dos elementos, como, por exemplo, dos integrantes do fluxo de caixa da empresa, representa uma parte importante do planejamento estratégico de uma entidade que poderia gerar, quem sabe, mais rapidez e economia nesse processo. A utilização dos métodos estatísticos mostra-se eficaz nessas situações, justificando sua utilização devido, principalmente, aos seguintes fatores:

1.2.2.1 Rapidez e precisão nas respostas aos problemas

O fator tempo, atualmente, é um recurso muito valorizado no meio corporativo. A cada dia, a velocidade das informações torna-se um diferencial importantíssimo, influenciando e, muitas vezes, determinando a continuidade de um negócio. Por outro lado, a ânsia pela velocidade pode acarretar o prejuízo da qualidade da informação. Tanto quanto rápidas, elas necessitam ser corretas, precisas. É a relação fundamental de tempestividade e precisão. A partir da utilização de métodos estatísticos adequados, corretamente escolhidos e modelados antes da aplicação prática, torna-se possível a obtenção de informações tempestivas e precisas.

1.2.2.2 Descentralização

O paradoxo da descentralização das informações em relação à evolução dos sistemas a partir de sua estruturação em bancos de dados — que centralizam o armazenamento das informações a serem processadas —, é uma realidade nas empresas que romperam seus limites regionais. A busca por processamentos descentralizados, mesmo a partir de dados centralizados, é uma verdade. Empresas com filiais, sucursais, *join-venture*, parceiras e afins beneficiam-se ao possibilitarem que determinadas operações sejam executadas de forma descentralizada. Isso contribui para a rapidez discutida no item imediatamente anterior. Nesse sentido, esse benefício refere-se à possibilidade de que pessoas, nem sempre detentoras de conhecimentos profundos em estatística, possam executar tarefas de análise no exato local onde se encontram. A descentralização pode contribuir para a dissipação do conhecimento de determinados processos de análise quantitativa, refletindo positivamente na cultura de controle da organização.

1.2.2.3 Maior possibilidade de testes antes de decisões importantes

Quanto custaria para uma empresa de fabricação de caixas-d'água de grandes proporções verificar a eficácia do concreto após este estar parcialmente aplicado e em processo de secagem? Certamente, seria um desastre econômico e financeiro.

Para esse tipo de procedimento, a aplicação dos conceitos da estatística torna-se eficaz para que, por meio dos procedimentos adequados, possam-se executar tantos testes quantos forem necessários, até que se tenha a garantia probabilisticamente calculada de que aquele material atingirá satisfatoriamente seus objetivos.

Os métodos estatísticos de análise por amostragem possibilitam a reaplicação de testes em fragmentos estatisticamente significantes, de tal forma que a repetição não comprometa o todo, permitindo a antecipação de possíveis problemas que uma má análise de qualidade poderia proporcionar.

1.2.2.4 Economia de recursos

Diante da realidade de que as empresas estão operando com recursos cada dia mais escassos — portanto, mais dispendiosos —, a sua economia, sejam tais recursos materiais, humanos ou financeiros, é imprescindível.

A pressão imposta pela competitividade e pelo "amadurecimento" da demanda, que impõe restrições às políticas de preços, praticamente obriga as empresas a buscar alternativas de contenção.

Recuperando o exemplo do tópico anterior da empresa de caixas-d'água, a possibilidade de, por meio de processo de amostragens, a empresa poder inferir opinião sobre um todo é, certamente, uma forma de se economizar tempo e dinheiro, além de se diminuir a utilização de recursos humanos necessários no caso de uma análise mais ampla para obter o mesmo resultado.

É importante ressaltar que a economia de recursos de qualquer ordem, como resultado da aplicação dos métodos estatísticos, representa algo que não pode interferir na segurança da análise em curso, ou seja, vale a pena a revisão constante da adequação dos procedimentos identificados, modelados e aplicados.

Seção 2 A informatização da estatística

Caro aluno, nos dias de hoje, para quase todos os processos que envolvem o trabalho com números, dados e informações em geral, foram desenvolvidos instrumentais ligados aos desenvolvimentos de softwares e informatização.

O mesmo ocorre com a Estatística: foram desenvolvidos softwares mais avançados, como Minitab, SPSS e Stata, e também outros recursos mais tradicionais, como os elaborados pela Microsoft no Excel.

Enfim, todos esses instrumentais tanto ampliaram quanto tornaram mais práticas as atividades de análise de dados, pois permitem realizar diversas medições estatísticas além da criação de gráficos, tomando por base os dados levantados em uma pesquisa.

Tendo em vista que para você, aluno, no ambiente corporativo, a ferramenta mais prontamente disponível é o Excel, vamos, então, estudar um pouco mais sobre essa ferramenta e suas aplicações estatísticas.

2.1 Noções de estatística com o uso do Excel

Com a informatização, a aplicação dos métodos estatísticos vem se tornando mais simples, rápida e eficaz. Focando os benefícios da aplicação desses métodos no meio corporativo, pode-se dizer que é possível, para as empresas, desenvolver inúmeros tratamentos de dados, bastando para isso dispor, por exemplo, de uma planilha eletrônica — como o Excel da Microsoft.

Questões para reflexão

Será que somente o uso da tecnologia resolve o problema em termos decisórios nas empresas? Estaria o profissional realmente ameaçado pelo computador no que se refere a análises e tomadas de decisão?

Existem outros softwares que possibilitam a aplicação via planilha eletrônica, mas o Excel ainda é o mais conhecido. Além das planilhas, algumas calculadoras, tanto financeiras quanto científicas, possuem funções estatísticas.

Apesar de as planilhas e calculadoras estarem cada dia mais evoluídas, ao menos outras duas formas de automatização são bastante eficazes: os suplementos para planilhas eletrônicas, que automatizam funções específicas, e os

softwares especializados. Certamente, de todas as opções, os softwares são os mais completos e robustos para as aplicações estatísticas.

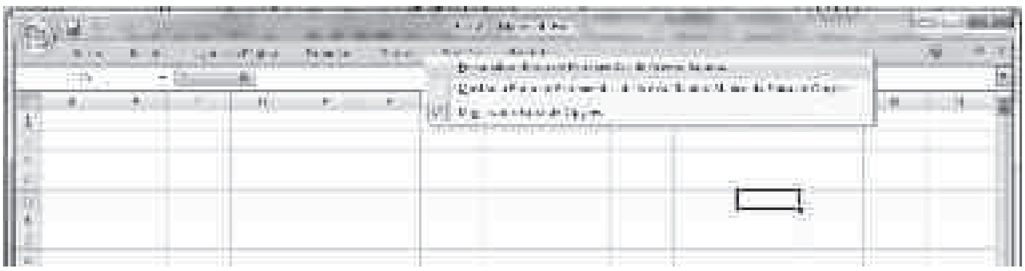
Com uma simples pesquisa na internet, é possível encontrar qualquer uma dessas formas de automatização das funções estatísticas. Existem, inclusive, aplicações sem custo para o usuário. Devido ao fato de que a maioria dos computadores, tanto os desktops quanto os notebooks, possui o sistema operacional Windows® com o pacote Office®, no qual consta a planilha Excel®, essa será a opção para o desenvolvimento dos exercícios presentes neste capítulo.

Para que se habilite o programa para as funções que serão trabalhadas ao longo do capítulo, algumas operações deverão ser desenvolvidas antecipadamente, conforme pode ser visto no Quadro 1.7. Nele, é possível observar o primeiro procedimento de preparação do software. O usuário deve habilitar alguns suplementos que são similares a uma biblioteca de funções que ficam desabilitadas na instalação normal — isso se deve, principalmente, à economia do uso de memória do computador.

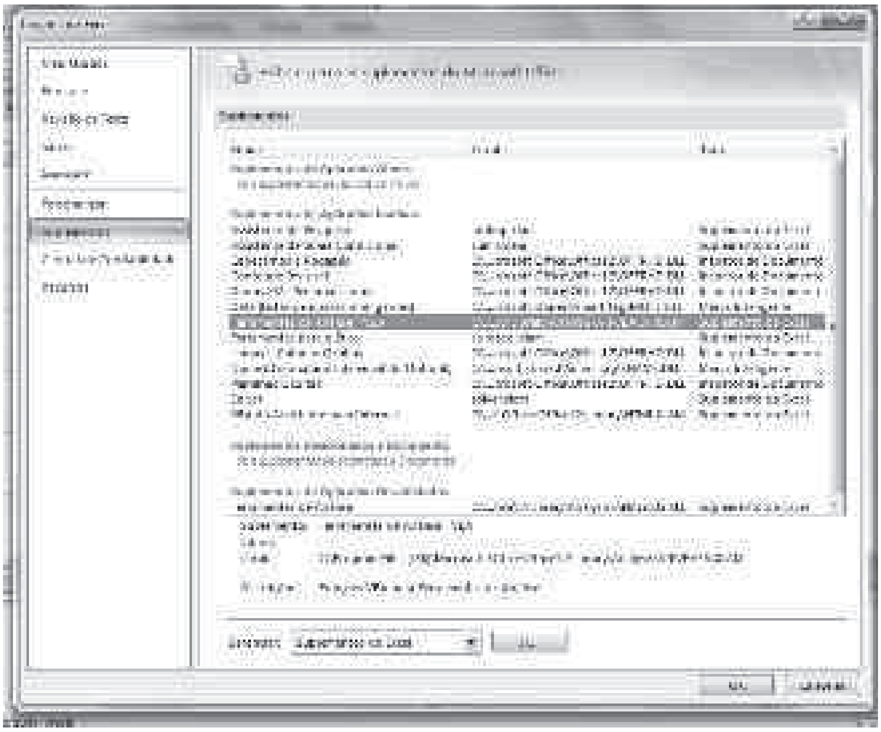
Para isso, o usuário deve clicar com o botão direito do mouse sobre a barra de ferramentas do Excel 2007 e marcar a opção "Personalizar Barra de Ferramentas de Acesso Rápido" (Parte A). Isso abrirá a janela "Opções" do Excel. Nesta, ele marcará a opção "Suplementos" e, nas opções de suplementos, marcará a opção "Ferramentas de Análise — VBA"; na opção abaixo onde aparece "Gerenciar", deverá escolher a opção "Suplementos do Excel" e clicar em "Ir" (Parte B). Será aberta outra janela "Suplementos", e ele deverá marcar as opções "Suplemento de Análise" e "Suplemento de Análise — VBA"(Parte C); ao clicar em OK, as ferramentas serão instaladas na barra de ferramentas quando você acessar a aba "Dados", ficando disponível do lado direito da tela em "Análise de Dados" (Parte D).

Quadro 1.7 Telas de habilitação de suplemento software Excel 2007

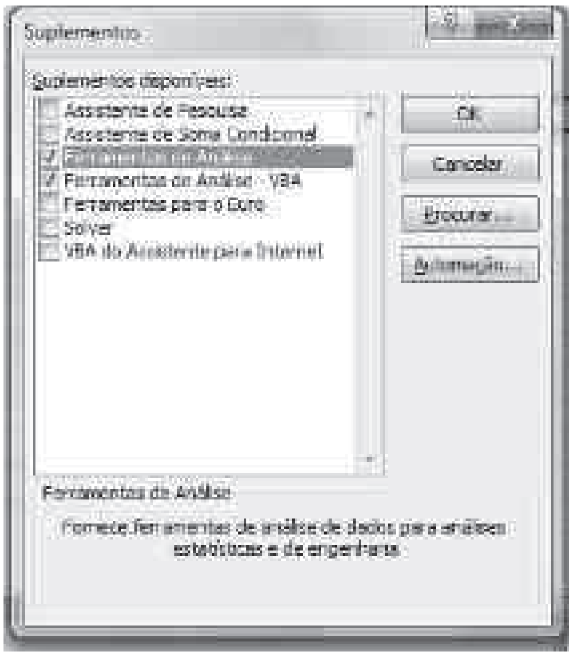
Parte A



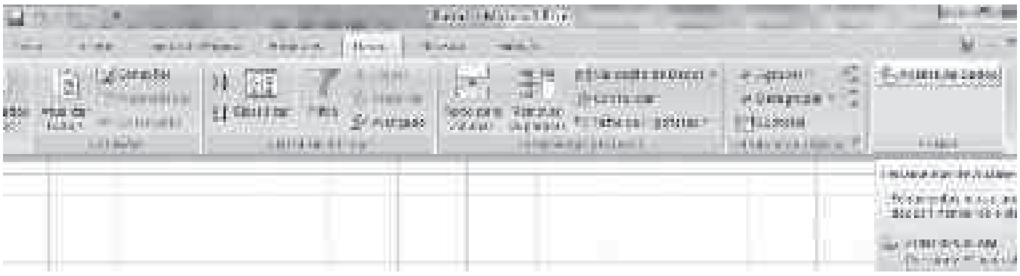
Parte B



Parte C



Parte D



Fonte: Dos autores (2014).

Vale lembrar que a forma de habilitação dos suplementos pode variar de acordo com a versão do software que se estiver utilizando. Nesse caso, aconselha-se a pesquisa nos manuais ou menus de ajuda do fabricante. As ferramentas habilitadas serão úteis para a automatização das análises estatísticas dos dados, incluindo a possibilidade de automatização do processo de amostragem, que será tratada em momento oportuno. Uma vez habilitados esses suplementos, o software disponibilizará as opções de análise de dados, que serão conhecidas à medida que houver a necessidade e que exercícios referentes a cada tópico sejam solucionados. A partir do uso da informática, é possível a agilização dos processos conforme foi demonstrado. Existem materiais bibliográficos específicos para orientá-lo sobre como utilizar esses recursos tecnológicos.

Para saber mais

Para estudar mais sobre informatização e estatística, você pode ler o Capítulo 1, "Introdução ao Excel", do livro **Estatística aplicada à administração usando o Excel**, do autor John L. Neufeld, publicado em 2003.

Pois bem. Agora que você compreende um pouco melhor o uso do Excel, podemos propor algumas atividades básicas. Vamos a elas.

Atividades de aprendizagem

1. Os dados abaixo representam o consumo de energia elétrica rural no estado do Paraná para os anos de 1999 a 2012. Observe os dados e faça o que se pede:

Ano	Energia elétrica rural — consumo (Mwh)
1999	1.098.565
2000	1.145.548
2001	1.153.589
2002	1.233.387
2003	1.268.096
2004	1.339.693
2005	1.409.379
2006	1.452.764
2007	1.545.217
2008	1.630.900
2009	1.706.287
2010	1.801.680
2011	1.899.936
2012	2.054.625

a) No Excel, copie os dados da tabela.
b) Crie gráficos: um de linha e outro de barra.

2. Os dados a seguir referem-se ao número de trabalhadores no setor agropecuário do Estado do Paraná para os anos de 2000 e 2010. Observe os dados e faça o que se pede:

Tamanho do estabelecimento de acordo com o número de empregados	Emprego total 2000	Emprego total 2010
ATE 4	303.111	380.709
DE 5 A 9	141.557	194.085
DE 10 A 19	121.890	173.309
DE 20 A 49	134.081	176.925
DE 50 A 99	83.093	119.945
DE 100 A 249	88.129	123.950
DE 250 A 499	53.661	72.883
DE 500 A 999	62.934	76.918
1000 OU MAIS	83.815	90.873
Total	1.072.271	1.409.597

a) Copie os dados para o Excel.
b) Crie uma terceira coluna para calcular quanto cresceu o emprego de 2000 para 2010.
c) Crie um gráfico de barras comparando os dois anos.

Seção 3 Organização e deflação de dados

Nesta seção, vamos nos ocupar em entender como se faz a organização de dados, isto é, para que nos utilizemos de dados, é necessário que, antes, sejam tomados alguns procedimentos prévios, pois são estes que vão fazer com que os dados levantados em uma pesquisa passem a fazer sentido ou permitam atingir um objetivo. Nesse sentido, cabe aqui uma pequena explanação sobre os passos para a realização de uma pesquisa experimental. Vamos estudá-los.

3.1 Sugestão de roteiro para as pesquisas descritiva e experimental

Escolha o assunto: devemos escolher um assunto significativo, relevante e adequado tanto aos interesses do pesquisador quanto aos interesses da comunidade acadêmica ou científica, de acordo com o nosso nível de formação e também com as condições do pesquisador.

Título da pesquisa: o título deve deixar claro qual o tema que está sendo trabalhado.

Delimitação do assunto: são muitos os assuntos que podem ser abordados; entretanto, às vezes eles são muito amplos. Assim, devemos selecionar um tópico para ser estudado de acordo com sua profundidade, de forma que seja viável de ser pesquisado. Devemos evitar temas amplos, pois eles podem tornar os trabalhos superficiais.

Objetivos: para os objetivos, devemos esclarecer o que pretendemos alcançar com a pesquisa.

Justificativa da escolha: nesse ponto, devemos explicitar as razões da preferência pelo assunto escolhido, ou seja, porque resolvemos estudar esse tema e, principalmente, qual é a sua importância em relação a outros temas.

Revisão da literatura: nessa etapa, devemos realizar uma pesquisa bibliográfica referente ao assunto e à questão delimitada. Esse estudo teórico tem como objetivo evidenciar os trabalhos já realizados a respeito do assunto, apresentar as informações sobre a situação atual do problema pesquisado e também explicitar as opiniões existentes. Essas informações são importantes para auxiliar o pesquisador nas tomadas de decisão.

Formulação do problema: o problema deve ser formulado de forma clara, precisa e objetiva e deve ser uma questão cuja solução viável possa ser alcançada pela pesquisa. Para elaborarmos um problema de forma clara, é necessária muita leitura e a revisão da literatura exposta no item anterior.

Enunciado da hipótese: a hipótese consiste em uma resposta antecipada ou uma explicação provisória do que o pesquisador enuncia. Em outras palavras, a hipótese é a solução que o pesquisador acredita. Ela deve ser colocada à prova e responder o problema levantado.

Definição operacional das variáveis: a hipótese vai orientar a execução da pesquisa. Dessa forma, os termos empregados na hipótese devem esclarecer, da melhor maneira possível, o que eles significam no contexto concreto e objetivo da pesquisa a ser realizada.

Amostragem: muitas vezes, ou melhor, na maioria delas, é difícil para o pesquisador trabalhar com a totalidade de seus elementos, ou seja, com toda a "população" ou "universo". Podemos compreender a população como um conjunto de pessoas, animais ou objetos que representam a totalidade de indivíduos que possuem as mesmas características definidas para um estudo. Geralmente, como não temos acesso a toda a população, podemos realizar a pesquisa com uma parte representativa da população, denominada "amostra", e não com a totalidade dos indivíduos. Dessa forma, podemos compreender a amostra como sendo uma parte da população selecionada de acordo com uma técnica de amostragem que garante sua representatividade.

Instrumentos de pesquisa: são os instrumentos utilizados pelo pesquisador para a coleta de dados, que podem ser bem variados: entrevista, questionário, formulário, medições etc. Quando a pesquisa realizada tem caráter experimental, então são descritos os instrumentos, materiais ou técnicas a serem usados.

Procedimentos: nessa etapa, é necessário realizar a descrição detalhada de todos os passos da coleta e registro dos dados: quem participou da pesquisa, onde foi realizada, quando, quanto tempo durou etc. Podemos, ainda, descrever as dificuldades encontradas no decorrer da pesquisa, as precauções, a supervisão e o controle.

Análise dos dados: após a coleta dos dados, é necessário realizar a análise dos mesmos. Os resultados da análise podem ser expostos em tabelas, de forma sintética, e submetidos ou não (de acordo com cada pesquisa) a um tratamento estatístico mais profundo em que todas as informações reunidas nos passos anteriores são comparadas entre si e analisadas.

Discussão dos resultados: nessa etapa, explicitamos a generalização dos resultados obtidos por meio da análise realizada. Nessa discussão, o pesquisador faz as inferências e generalizações pertinentes com base nos resultados alcançados. Os resultados obtidos também são discutidos e comparados com afirmações e posições de outros autores investigados pelo pesquisador.

Conclusão: na conclusão, devemos apresentar um resumo dos resultados mais significativos e relevantes da pesquisa e indicar se esses resultados auxiliam na comprovação ou rejeição da hipótese de estudo já apresentada. Também é o momento do pesquisador indicar quais os possíveis desdobramentos de sua pesquisa, ou seja, quais os aspectos que podem ser estudados e aprofundados em outros estudos.

Bibliografia: o pesquisador deve citar, de acordo com normas próprias, as referências bibliográficas que serviram de embasamento para o seu estudo teórico.

Pois bem. Como vimos, os passos da pesquisa são importantes porque nos direcionam para atingir o objetivo e também preparam o caminho para que se possa escrever o relatório dessa pesquisa. Assim, os itens análise de dados, discussão de resultados, conclusão e bibliografia fazem parte de qualquer relatório de pesquisa, assim como os itens anteriores de título, introdução, justificativa, problema, revisão de literatura etc. Mas um item das etapas de pesquisa que é indispensável é a fase de análise de dados, cuja pré-condição é que estes tenham sido agrupados ou modelados de alguma forma. Assim, precisamos ver como os dados podem ser organizados.

É o que vamos estudar no próximo item.

3.2 Organizando e modelando os dados

Pois bem. Após a realização do planejamento da pesquisa, é necessário definir de que maneira vamos levantar as informações necessárias, quais são os dados que devemos coletar, que tipo de amostra devemos usar, qual é o cronograma das atividades etc.

A coleta de dados se caracteriza por ser prática ou operacional. Nela, fazemos o registro sistemático de dados obtidos por meio da coleta, com um objetivo pré-estabelecido. Podemos ter alguns tipos de dados:

- **dados primários:** quando o pesquisador utiliza e publica os dados que ele mesmo ou sua organização coletou, como, por exemplo, os dados do censo demográfico do IBGE.
- **dados secundários:** são dados publicados por outro tipo de organização, como, por exemplo, quando um jornal publica estatísticas obtidas pelo IBGE. Uma observação importante é que é sempre mais seguro utilizar fontes primárias. O uso de fontes secundárias pode provocar o risco de erros de transcrição. Nesse caso, podemos fazer uma **coleta direta,** isto é**,**

obter os dados diretamente da fonte, como, por exemplo, uma empresa que realiza uma pesquisa para observar qual a preferência dos consumidores em relação a sua marca de produtos.

Após a coleta de dados, podemos entrar na fase de **apuração dos dados**. De posse dos dados obtidos, os mesmos são contados e agrupados por meio da tabulação dos dados. Essa fase também é chamada de tabulação de dados; ela indica que os dados estão sendo agrupados de alguma forma para que possam ser apresentados e utilizados na pesquisa. Na **apresentação dos dados,** pode-se optar por uma apresentação **tabular,** na qual os dados são apresentados em sua notação numérica, organizados em linhas e colunas ou, ainda, na forma **gráfica** em que os dados são apresentados de forma geométrica, que permite uma visualização rápida e uma compreensão clara do fenômeno estudado.

3.2.1 Modelando os dados

À medida que uma determinada situação exija um tratamento estatístico, repetindo-se costumeiramente, será importante que haja uma customização desse processo a ponto de automatizá-lo e torná-lo padronizado.

Todo o processo de análise estatística pressupõe que haja o respeito à sequência apresentada na seção anterior, ou seja: obtenção dos dados; descrição, classificação e apresentação dos dados e conclusões sobre os dados.

Você verá que essas etapas são representadas por metodologias capazes de dar sentido e garantir a inteligibilidade dos dados bem como das análises. Metodologias, como os números índices, tabulação, elaboração de gráficos etc., compõem o conjunto de elementos que tornam o tratamento estatístico algo consistente e, sobretudo, didático.

A análise prévia da necessidade informacional que o usuário pretende a partir da situação objeto é o que determinará os tipos de ferramentais estatísticos a serem utilizados e, num segundo momento, automatizados, devendo ser sugerida a criação de um **modelo** para ser utilizado quantas vezes se **desejar executar a análise**. A repetição poderá significar, inclusive, aperfeiçoamento dos modelos usados à medida que se permita a contribuição prática e teórica dos envolvidos.

Modelo é uma “representação simplificada da realidade, estruturada de forma tal que permita compreender o funcionamento total dessa realidade ou fenômeno” (MATOS, 1995, p. 20).

Em resumo, denomina-se modelagem estatística o processo de identificação dos procedimentos adequados para se efetuar o tratamento dos dados de forma a atender a necessidade de informação do analista. Para que haja confiabilidade na aplicação de um modelo estatístico, alguns princípios devem ser respeitados:

- Deve possuir uma teoria de sustentação;
- Deve ser "inteligível" para seus usuários;
- Deve possuir uma metodologia adequada em termos de complexidade e extensão;
- Deve possibilitar a terceiros sua replicação como forma de confirmação ou refutação das respostas apresentadas.

Questões para reflexão

Você acha que um cálculo estatístico que seja feito apenas para pessoas formadas em estatística entenderem atinge seu objetivo? Por quê?

A resposta à questão proposta talvez o ajude a entender a importância da modelagem, já que ela proporciona inteligibilidade às aplicações estatísticas. A modelagem pressupõe, em segunda análise, a possibilidade de automatização, o que agilizará futuros tratamentos e permitirá que os cálculos sejam feitos por outros profissionais, proporcionando a socialização da informação.

Ela contribui, então, para o tratamento dos dados e para o acesso às análises ao longo do tempo. Existem alguns tipos de análises feitas a partir do tratamento dos dados. Geralmente, eles variam em relação aos aspectos: profundidade da análise e nível de estudo da relação entre as variáveis envolvidas.

Sob o aspecto da profundidade, quanto mais complexa for a análise, mais desejável se torna a aplicação de modelos. Isso também ocorre no que se refere à busca do conhecimento sobre o relacionamento entre variáveis. Uma vez estabelecida essa relação, o modelo poderá contribuir para a automação de previsões, por exemplo.

3.2.1.1 Tabulação

Tabulação é a fase que ocorre após a coleta dos dados para análise e pressupõe que se saiba antecipadamente seus objetivos. A antecipação dos objetivos

de análise é o que evitará, por exemplo, trabalhos desnecessários e perda de tempo e de recursos financeiros ao longo de sua pesquisa ou análise gerencial dentro da empresa.

Antes de serem tabulados, os dados coletados representam meros números ou aspectos qualitativos soltos e de difícil interpretação. Ao se tabular, cumpre-se um primeiro passo para tornar esses dados mais inteligíveis e analisáveis.

A expressão visual de um dado ou informação é certamente uma das formas mais eficazes, em termos didáticos, para que sejam captadas as nuances contidas neles. Essa percepção visual melhora a compreensão do analista e dos usuários do objeto em análise, além de representar uma forma bastante eficaz de organização e recuperação de informações relevantes.

O que você pode entender, então, por tabulação? Lembre-se: a tabulação é a **sistematização dos dados** coletados para a análise, de forma a torná-los organizados e inteligíveis, possibilitando a execução do tratamento necessário para que sejam atingidos os objetivos previamente traçados.

Atividades de aprendizagem

1. Veja no exemplo do quadro a seguir que, após a tabulação, já é possível uma análise, mesmo que superficial, dos dados. Isso se dá devido a sua organização de acordo com os objetivos. Que conclusão você poderia tirar dos dados apresentados no Quadro 1.8?

Quadro 1.8 **Dados hipotéticos**

SEQUÊNCIA DA OBSERVAÇÃO	NÚMERO DE TRAVAMENTOS DO EQUIPAMENTO	CUSTO GERADO ($)
1	8	1600,00
2	12	2400,00
3	6	1200,00
4	7	1400,00
5	5	1000,00
6	8	1600,00
7	15	3000,00
8	20	4000,00
9	12	2400,00
10	8	1600,00

2. Quando os valores de uma observação se repetem, é possível fazer uma tabulação agrupando os dados de acordo com sua frequência, isto é, agrupando o número de vezes que aparece. Vamos imaginar que uma empresa X realizou uma pesquisa sobre a satisfação dos clientes em relação ao atendimento recebido e, para poder quantificar essas informações, solicitou que eles associassem notas a graus de qualidade do atendimento recebido. Assim, quando o atendimento era ótimo, a nota era 5, quando era bom, 3, quando regular, 1 e, para ruim, 0. Ao conferir as notas atribuídas à empresa, deparou-se com os seguintes números: 0, 3, 1, 1, 3, 5, 1, 0, 0, 1, 3, 5, 3, 1, 0, 5, 3,3. Observando esses dados, complete a tabela.

Tabela 1.1 **Satisfação dos clientes da empresa X**

Conceito	Nota	Frequência
Ótimo	5	
Bom	3	
Regular	1	
Ruim	0	

Com base nos dados tabulados na tabela, como foi classificado o atendimento da empresa de acordo com as notas atribuídas pelos clientes?

3.3 Deflacionando dados: números-índices

Os números-índices são medidas usadas para comparar grupos de variáveis que possuem algum relacionamento entre si, de tal forma que seja possível a obtenção de um quadro organizado dos comportamentos das variáveis dentro desse grupo selecionado, como, por exemplo: variação de preço do produto A num determinado espaço de tempo, ou uma comparação do produto A com o produto B em termos de preço em um determinado período de tempo. Também chamado de deflacionamento de dados.

Pode ser definido como um instrumento estatístico utilizado por várias áreas, como administração, economia, contabilidade, engenharia etc., para fazer comparações de alterações em variáveis específicas num dado espaço de tempo. Geralmente,tratam-se de variáveis como preço, quantidades, volume

de produção etc. de um elemento qualquer ou de vários elementos tomados para comparação.

Se você está diante de uma análise, por exemplo, de capacidade de poder de compra envolvendo o salário da população, certamente utilizará números-índices como forma de comparação.

Os números-índices utilizados em estatística contribuem para as análises, principalmente comparativas, de períodos. O índice de preços ao consumidor (IPC), por exemplo, contribui para que se possa fazer uma análise histórica da variação de preços de vários itens da cesta básica de consumo da população. Como se poderia fazer essa comparação sem a padronização dos itens que compõem o índice e sem trazê-los para uma mesma unidade de análise, nesse caso podendo ser a variação percentual de um período para outro?

Os números-índices, ao se referirem a um único elemento sendo comparado em períodos distintos, são denominados números-índices simples. Ao envolver mais de um elemento, os números-índices são denominados compostos.

É comum a aplicação desse tipo de metodologia em cálculos econômicos, como pode-se verificar na citação a seguir:

> A teoria dos números-índices possui duas ramificações no que diz respeito ao número de períodos ou de regiões a serem comparados: números-índices bilaterais, quando o número de unidades econômicas é igual a dois, e números-índices multilaterais, quando esse número é superior a dois (AZZONI; CARMO; MENEZES, 2003, p. 4).

É possível verificar, portanto, que existem outras classificações possíveis (bilaterais e multilaterais) no que se refere aos números-índices; porém, basta para os objetivos desta unidade o entendimento sobre seu conceito e aplicabilidade. Para verificar uma aplicação de números-índices, recorremos a um exemplo:

CASO I: Correção Monetária

Em setembro de 2007, uma pessoa realizou um empréstimo de R$ 90,00 em um banco. Passados três meses, essa pessoa resolveu suspender os pagamentos, apresentando na justiça sua inconformidade com o valor a ser pago. A disputa judicial prosseguiu 5 anos. Ao final, chegou-se a um acordo em que o cidadão deveria pagar a dívida corrigida pelo IGP-M.

Solução

Nesse caso, a dívida está sujeita à correção monetária como forma de atualizar o seu valor. A forma de realizar esse cálculo é dividir o indexador do período atual pelo indexador do período base.

Então, se em setembro de 2007 (período base), o IGP-M foi 871,11, e em setembro de 2007 (período atual), o mesmo índice chegou a 1.273,27, a correção monetária poderá ser realizada multiplicando-se o valor base ou inicial, R$ 90,00, pelo resultado da divisão dos indexadores **(1,4616)**, o que dará um valor corrigido de **R$ 131,54**.

Fórmula

Valor corrigido = (índice atual / índice base) x valor base

CASO II: Deflacionamento

Podemos imaginar, partindo da situação anterior, que a sentença tivesse ordenado que fosse pago apenas o valor devido, ou seja, R$ 90,00. Tendo em vista que durante todo esse tempo houve inflação, então gostaríamos de saber qual foi o valor real, isso é, o valor pago, descontada a inflação.

Solução

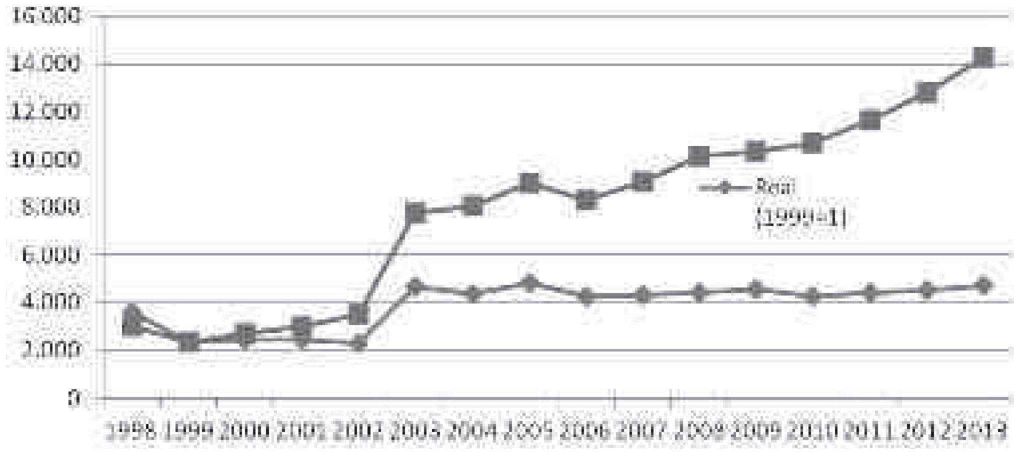
Nesse caso, estaremos realizando uma operação inversa à do caso I. Assim, a fim de eliminar o efeito da inflação, dividimos o valor monetário do período atual pelo valor obtido da divisão entre o indexador do período atual pelo indexador do período. Tomamos R$ 90,00 e dividimos por 1,4646 e teremos o valor de R$ 61,45. Esse valor final representa que, caso o cliente pagasse R$ 90,00 hoje, contraídos há cinco anos, o valor verdadeiro do pagamento seria apenas R$ 61,45. Na prática, seria o mesmo efeito de ter contraído R$ 61,45 em empréstimo e pago R$ 90,00, considerando a inflação.

Fórmula

Valor deflacionado = valor base / (índice atual / índice base**)**

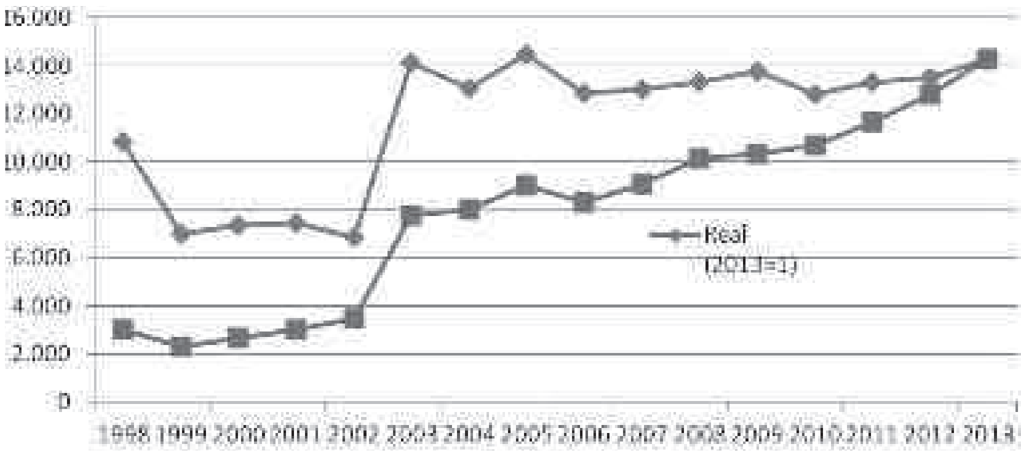
Nos gráficos a seguir, você pode visualizar o que ocorre quando os valores nominais (linha com marcação quadrado) são deflacionados (linha com marcação losango). No Gráfico 1.1, os valores foram deflacionados. Isso significa que todos os valores nominais foram transformados em valores reais para o ano de 1999=100. No Gráfico 1.2, o valores nominais foram inflacionados. Isso significa que os valores foram transformados em valores reais para o ano de 2013 = 100.

Gráfico 1.1 **Exemplo de valores reais deflacionados**



Fonte: Dos autores (2014).

Gráfico 1.2 **Exemplo de valores reais inflacionados**



Fonte: Dos autores (2014).

3.4 Um pouco sobre inferência estatística

É a aplicação de técnicas estatísticas segundo determinados princípios e critérios válidos, de forma a permitir a extrapolação das análises feitas de uma amostra para o todo. Implica, geralmente, a conclusão sobre causa e efeito de determinadas variáveis.

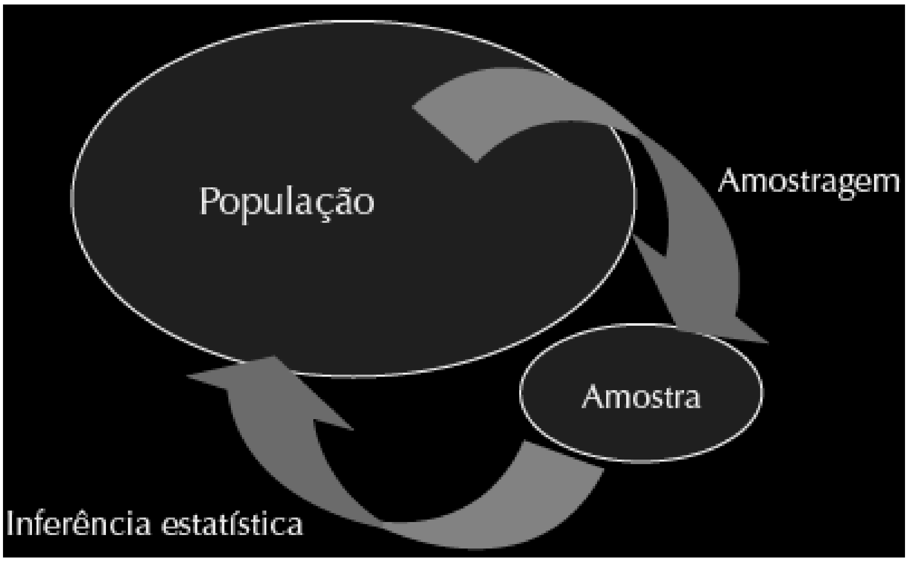
A inferência pressupõe que os resultados obtidos na amostra sejam passíveis de ser extrapolados para toda a população em estudo. Portanto, exige que sejam utilizadas técnicas científicas para a determinação da amostra que servirá como objeto de análise do todo.

Não há inferência estatística sem o pressuposto da adequação dos métodos utilizados, ou seja, novamente se destaca a importância de que os modelos sejam construídos a partir de um minucioso estudo das características do objeto em estudo, de tal forma que as técnicas a serem utilizadas possam ser identificadas adequadamente. Um erro nesse processo implica outro na inferência.

O processo inferencial também pode ser classificado como indutivo, uma vez que parte de uma amostra para o entendimento do todo — o que possibilita economia de recursos para as empresas, conforme já mencionado anteriormente. O processo indutivo é menos agressivo para a população analisada, principalmente no caso de esta, por exemplo, referir-se às análises materiais de qualidade.

Como seria a remoção de todos os tijolos de uma olaria para um teste de qualidade em laboratório? O Quadro 1.9 demonstra a relação da estatística inferencial com a amostra e a população estudada.

Quadro 1.9 **Relação população amostragem e inferência**



Fonte: Dos autores (2014).

Para que você visualize o conteúdo exposto nessa unidade de forma prática, é apresentado na sequência um exemplo simplificado no qual se retoma cada uma das subseções discutidas.

Imagine a seguinte situação:

A empresa Analisando Ltda. pretende saber o grau de satisfação de seus clientes. Ela dispõe para tanto de uma população, é claro. Qual seria essa população?

Para responder a essa questão, o dono da empresa chamou o senhor Mente Aberta, um funcionário de seu primeiro escalão e que recentemente fez um curso sobre estatística aplicada.

O dono da empresa apresentou ao senhor Mente Aberta sua primeira questão — qual seria a população a ser analisada? —, que foi respondida da seguinte forma:

A POPULAÇÃO: *o conjunto de todos os clientes que já consumiram os produtos da empresa Analisando Ltda*. Veja que a população foi claramente definida pelo funcionário, inclusive em seus detalhes. Isso é muito importante quando se pretende fazer uma análise estatística.

Agora é o senhor Mente Aberta que dialoga com o dono da empresa:

O primeiro passo para que consigamos chegar a um bom resultado, o senhor já deu, ou seja, definiu o que deseja a partir da análise da população.

"Saber o grau de satisfação de nossos consumidores". Para nosso segundo passo, é necessário, então, que façamos uma amostragem, pois seria impossível entrevistarmos todos os nossos clientes. São muitos. Eu sugiro, então, que, do total de nossa população, analisemos 20%, pois, pelos meus cálculos, é essa a quantidade que nos garantirá significância estatística para a amostra.

Como você viu, o senhor Mente Aberta está preocupado com outro fator muitíssimo importante para a estatística aplicada, ou seja, a significância estatística da amostra.

Bem! Para chegar, então, à amostra de 20%, o senhor Mente Aberta solicitou ao departamento de vendas a relação dos clientes que a empresa atende regularmente. Esse relatório contou com 150 clientes. Então, qual seria a amostra?

AMOSTRA: *a amostra, segundo determinou o funcionário da empresa, é calculada, portanto, a partir da aplicação de 20% sobre a população de 150 clientes. Dessa forma, a amostra será de 30 clientes.*

A partir desse segundo passo importante, já é possível, então, definir-se todos os elementos fundamentais para a aplicação da análise estatística. Veja o Quadro 1.10 e relacione com os conceitos vistos na unidade anterior.

Quadro 1.10 **Definições para os parâmetros estatísticos no exemplo**

ELEMENTO	IDENTIFICAÇÃO
População ou universo	Todos os clientes que compraram na empresa e que se pretendem pesquisar. 150 clientes.
Elementos ou unidades	Cada um dos clientes representa um elemento ou unidade.
Amostra	A porção de clientes que será entrevistada e definida no exemplo como 20% da população, o que equivale a 30 clientes.
Variável	A característica específica que se pretende analisar da população ou da amostra. No caso, o índice de satisfação.

Fonte: Dos autores (2014).

Antes de avançarmos no exemplo, é importante destacar que o próprio termo "índice de satisfação" remete a um outro conceito estudado, ou seja, números-índices. Embora, conforme já foi mencionado, esse tipo de parâmetro seja mais utilizado na área econômica — na avaliação de preços e renda —, não podemos ignorar que, a partir do momento em que a empresa Analisando Ltda. começar a armazenar esse índice e fizer as comparações dele com ele mesmo em períodos distintos, ela estará se utilizando de um índice padrão comparativo. Essa é não só a definição simplificada, como também a aplicação de um número-índice importantíssimo.

Prosseguindo agora com o exemplo, passamos novamente a palavra para o dono da empresa, que chama mais um funcionário. O senhor Curiosinho é o responsável por coletar os dados dos clientes, e o proprietário lhe dá a seguinte incumbência:

Preciso que você entreviste 30 clientes e pergunte, numa escala de 5 a 10, o nível de satisfação deles em relação ao produto que compraram aqui em nossa empresa.

Antes de o senhor Curiosinho sequer sair para cumprir a incumbência dada pelo patrão, foi alertado pelo senhor Mente Aberta de que, ao selecionar os clientes para a entrevista, deverá fazê-lo de forma aleatória para que não haja viés no processo de seleção que comprometa toda a análise.

Questões para reflexão

Você sabe o que é uma amostra ou variável aleatória? Pesquise e descubra.

Voltando agora para o exemplo, depois de algumas horas, o senhor Curiosinho trouxe os dados anotados nas fichas de entrevistas e entregou-os ao dono da empresa. O dono agradeceu e fez a seguinte pergunta:

Senhor Mente Aberta, o que eu faço com esse "monte de dados"? Não me dizem nada.

Sem "pestanejar", o funcionário de primeiro escalão explicou o procedimento para que os dados ali apresentados se tornassem mais inteligíveis.

Devemos primeiramente tabulá-los, ou seja, organizá-los de tal forma que se tornem inteligíveis e já nos permitam alguma análise, mesmo que superficial, sobre esses dados.

E foi o que ele fez: tabulou os dados apresentando como resultado o quadro seguinte, respeitando a ordem na qual foi feita a entrevista como parâmetro de classificação.

Parece ficar evidente que o grau de satisfação dos clientes é razoavelmente bom. Basta observar que, aparentemente, a maioria das notas está acima de 7,0, o que indica um bom nível de satisfação.

A partir da tabulação, já foi possível ter uma visão preliminar dos dados. É claro que isso é possível em dados com menos variáveis, como no exemplo, e também quando o comportamento desses dados é perceptível a partir da simples observação.

Mas para se chegar a uma conclusão mais aprimorada, basta agora calcular algum parâmetro que ajude o dono da empresa a conhecer o índice de satisfação de seus clientes.

Para isso, o senhor Mente Aberta explica a seu patrão que é necessário algum critério para a análise. Um parâmetro que seja utilizado para definir os índices ou conceitos esperados: o que significa uma boa satisfação e o que significa um baixo nível de satisfação para a empresa.

Ele rapidamente sugeriu o seguinte:

Vamos considerar que notas entre 5 e 6 equivalham a "pouco satisfeito". Que notas entre 6,5 e 8,5 signifiquem "satisfeito" e que notas acima de 8,5 signifiquem "muito satisfeito". Dessa forma, podemos fazer um gráfico que nos permita visualizar como está o índice de satisfação de nossos clientes.

Quadro 1.11 Dados tabulados da pesquisa de satisfação

SEQUÊNCIA (CLIENTE)	NOTA ATRIBUÍDA AO PRODUTO
1	7,5
2	8,0
3	6,5
4	9,0
5	5,8
6	8,7
7	9,6
8	8,3
9	6,2
10	5,5
11	9,9
12	10,0
13	7,8
14	8,2
15	9,8
16	8,3
17	8,7
18	6,7
19	8,0
20	7,5
21	8,0
22	6,5
23	9,0
24	10,0
25	5,7
26	10,0
27	9,1
28	8,3
29	6,8
30	9,8

Fonte: Dos autores (2014).

O segundo passo, então, foi separar as notas dos clientes segundo cada um desses grupos. Além disso, resolveu-se tirar uma média das notas dentro de cada grupo e uma média geral para as notas da amostra.

Questões para reflexão

Você sabe o que é uma média?

Para saber mais

A construção de um índice é importante porque permite comparar valores de dois períodos distintos, por isso é uma ferramenta útil para as empresas. Elas podem criar seus próprios números-índices observando, por exemplo, a evolução nos preços e nas vendas de seus produtos. Em seu significado genérico, um número-índice consiste em uma média de variações relativas. Se as variações medidas são as correspondentes aos preços, um número-índice de preços deve ser construído, o mesmo acontecendo com outras variáveis, como taxas de câmbio, taxas de juros, salários etc. Se as variações medidas são as correspondentes às quantidades, um número-índice específico deve ser construído: por exemplo, o correspondente ao *quantum* da produção industrial, da agrícola, das exportações, das importações etc.

Fique ligado!

Nesta unidade, foram apresentados os conceitos básicos necessários à compreensão da estatística. Entre os tópicos discutidos, estão origens, objeto e objetivos da estatística, que destacam como se iniciaram as primeiras tentativas de compreensão dos dados a partir de ferramentas metodologicamente voltadas ao que hoje se denomina estatística. Destaca alguns dos principais termos utilizados na área da estatística, como os conceitos de população, unidade, amostra e variável. Além de ressaltar as vantagens da utilização das metodologias estatísticas, apresenta de forma contundente o papel da informatização como meio para sua socialização. Ademais, fez-se uma breve exposição sobre estatística descritiva, exploratória e inferência estatística.

Para concluir o estudo da unidade

Caro aluno, você deve estar ciente de que a Estatística é um dos mais importantes instrumentais de trabalho na gestão de negócios. É através dela que dados, informações, números tomam formas mais definidas, tornando-se, então, possíveis de serem utilizados nos processos de análise de negócios e tomadas de decisão. Assim, nesta primeira unidade, você tomou contato com os conceitos básicos. Para aperfeiçoar seus conhecimentos, continue estudando as demais unidades. Você deve, também, acessar nossa Biblioteca Digital e verificar que lá estão disponíveis outros livros com os quais você poderá aprofundar seus estudos. Aproveite esse processo de estudos, anote suas dúvidas e as envie ao seu tutor eletrônico e professor para que eles possam auxiliá-lo e, não esqueça... continue estudando!

Atividades de aprendizagem da unidade

1. Responda às questões:

a) Qual a importância da estatística para as empresas e para a academia?

b) Sintetize os conceitos de: população, unidade, amostra e variável, exemplificando cada um desses termos.

c) Quais os benefícios da utilização da estatística no cotidiano de uma empresa?

d) No que a informatização pode contribuir para o processo de análise de dados?

e) Quais são as três fases da análise estatística?

2. Classifique, de acordo com os enunciados, os tipos de amostragem:

a) Um candidato a vereador faz uma pesquisa de intenção de voto na sua região de residência.

b) Uma empresa realiza uma pesquisa com pessoas que estão circulando em uma praça principal de uma cidade.

c) Um representante comercial da fábrica de pão visita várias lanchonetes de uma cidade, pesquisando a preferência dos consumidores.

Referências

ARAMAN, E. M. de O. SAMPAIO, H. R. **Estatística**: marketing. São Paulo: Pearson Prentice Hall, 2009.

AZZONI, C. R.; CARMO, H. E. do; MENEZES, T. Comparações da paridade do poder de compraentre cidades: aspectos metodológicos e aplicação aocaso brasileiro. **Pesquisa e planejamento econômico**. São Paulo: IPEA, v. 33, n. 1, 2003.

FERNANDES, Flávio César F. **A pesquisa em gestão da produção**: evolução e tendências. Rio de Janeiro: Enegep, 1999. 1 CD-ROM.

GARCIA, Regis. **Estatística**: administração III. São Paulo: Pearson Prentice hall, 2010.

GUJARATI, D. **Econometria básica**. Rio de Janeiro: Elsevier, 2006.

LARSON, Ron; FARBER, Betsy. **Estatística aplicada**. 2. ed. São Paulo: Pearson, 2003.

MATOS, O. C. de. **Econometria básica:** teoria e aplicações. São Paulo: Atlas, 1995.

MEMÓRIA, José Maria Pompeu. **Breve história da estatística**. Brasília, DF: Embrapa Informação Tecnológica, 2004. Disponível em: <http://www.im.ufrj.br/~lpbraga/prob1/historia_estatistica.pdf>. Acesso em: 12 mar.2014

MIURA, D. **Treinamento permanente em Excel**. Disponível em: <http://www.expresstraining.com.br/index.php?page=article&id=188>. Acesso em: 15 mar. 2009.

NEUFELD, J. L. **Estatística aplicada à administração:** usando o Excel. São Paulo: Pearson, 2002.

SPIEGEL, M. R. **Estatística:** resumo da teoria, 875 problemas resolvidos, 619 problemas propostos. Tradução de Pedro Cosentino. São Paulo: McGraw-Hill do Brasil, 1977.

STOCK, J. H.; WATSON, M. W. **Econometria**. Tradução de Monica Rosemberg. São Paulo: Addison-Wesley, 2004.

Unidade 2

Índices e análises estatísticas

Marcelo Caldeira Viegas
Helenara Regina Sampaio
Regis Garcia
André Marcelo Santos de Souza
Kiliano Gesser
Márcia Vilma Depiné Dalpiaz

Objetivos de aprendizagem: Nesta unidade, você vai ser levado a compreender conceitos relacionados às medidas descritivas de dados, análise exploratória e descritiva e tabulação e análise de dados em gráficos, conceitos estes de extrema importância para a área de administração e que poderão ser utilizados pelos gestores nas tomadas de decisão.

Seção 1: **Medidas descritivas**

Nesta seção, serão apresentadas as medidas estatísticas que indicam medidas de tendência posição — dentre elas, os conceitos de média, moda e mediana. Também estudaremos os conceitos relacionados às medidas de dispersão.

Seção 2: **Análise exploratória e descritiva**

Nesta seção, serão estudadas as técnicas de exploração de dados estatísticos, em que serão abordadas de forma detalhada as fases do método estatístico e a organização de dados estatísticos.

Seção 3: **Tabulação e análise de dados em gráficos**

Nesta seção, serão apresentados de forma detalhada os conceitos de tabulação e análise de dados em forma de gráficos e tabelas de acordo com as normas técnicas citadas pela Associação Brasileira de Normas Técnicas (ABNT). Também serão dados vários exemplos de elaboração de gráficos utilizando o software Excel® da Microsoft.

Introdução ao estudo

É necessário ter em mente que a estatística é uma ferramenta essencial para o gestor ou administrador para as respostas dos "porquês" de seus problemas, que podem ser explicados por uma análise de dados. Para ela ser bem usada, é necessário conhecer os seus fundamentos e princípios e, acima de tudo, que o gestor desenvolva um espírito crítico e jamais deixe de pensar, pois é fácil mentir usando a estatística, o difícil é falar a verdade sem usá-la. Diariamente, somos expostos a uma grande quantidade de informações numéricas, semelhantes às relatadas. Dependendo das situações, ora somos consumidores dessas informações, ora precisamos produzi-las. Assim, necessitamos de capacitação para compreendermos informações numéricas produzidas por outros, bem como nos habilitarmos a construí-las. O emprego dos procedimentos, técnicas e métodos estatísticos é fundamental para auxiliar-nos na execução dessas tarefas.

Nesta unidade, que está dividida em três seções, serão abordados conceitos relacionados às medidas descritivas, à análise exploratória e descritiva e aos conceitos de tabulação e análise de dados em gráficos. Na Seção 1, abordaremos os conceitos relacionados às medidas descritivas, em que serão definidos alguns conceitos básicos da estatística descritiva. As medidas descritivas básicas mais importantes são as de posição e as de variabilidade, ou também conhecidas por dispersão. Nessa seção, serão apresentadas as medidas estatísticas que indicam medidas de tendência posição — dentre elas, os conceitos de média, moda e mediana. Também estudaremos os conceitos relacionados às medidas de dispersão, onde serão apresentados os conceitos de amplitude total, desvio médio, desvio-padrão (amostral e populacional), variância (amostral e populacional) e coeficiente de variação (CV). Na Seção 2, trabalharemos com os conceitos relacionados à análise exploratória e descritiva de dados, em que serão estudadas as técnicas de exploração de dados e abordadas de forma detalhada a organização de dados estatísticos e as fases do método estatístico. Na Seção 3, serão apresentados os conceitos de tabulação e análise de dados em forma de gráficos e tabelas de acordo com as normas técnicas citadas pela Associação Brasileira de Normas Técnicas (ABNT).

Seção 1 Medidas descritivas

A estatística descritiva está presente em praticamente todas as áreas do conhecimento humano, seja ajudando nos estudos e análises de dados, na busca de informações relevantes ou como elemento de comprovação científica de fenômenos relevantes. Conhecer e dominar alguns de seus métodos coloca o profissional em situação de vantagem no que se refere à competitividade e empregabilidade, além de proporcionar às organizações que dispõem desses profissionais meios de igualmente se tornarem mais competitivas. “O que é a estatística?” Vamos desvendar essa ciência? Então vamos lá! A estatística descritiva é a etapa inicial da análise utilizada para descrever e resumir os dados. A disponibilidade de uma grande quantidade de dados e de métodos computacionais muito eficientes faz com que essa área da estatística se torne essencial. As medidas descritivas têm por objetivo descrever um conjunto de dados de forma organizada e compacta por meio de suas funções básicas — dentre elas, destacam-se as medidas de posição e as de dispersão. O importante é destacar, neste momento, que esses cálculos e conclusões não podem ser extrapolados para a população e restringem-se apenas para a amostra.

A estatística descritiva pode ser definida como um conjunto de técnicas destinadas a descrever e resumir dados a fim de que possamos tirar conclusões a respeito de características de interesse. Em geral, utilizamos a estatística descritiva na etapa inicial da análise quando tomamos contato com os dados pela primeira vez. Objetivando tirar conclusões de modo informal e direto, a maneira mais simples seria a observação dos valores colhidos.

Entretanto, ao nos depararmos com uma grande massa de dados, percebemos, imediatamente, que a tarefa pode não ser simples. Para tentar retirar dos dados informações a respeito do fenômeno sob estudo, é preciso aplicar algumas técnicas que nos permitam simplificar a informação daquele particular conjunto de valores. A finalidade da estatística descritiva é tornar as coisas mais fáceis de entender, de relatar e discutir.

A estatística descritiva implica em descrever aquilo que os dados indicam após tabulados, convertidos em tabelas e gráficos. Como o próprio nome sugere, sua função é descrever aquilo que estamos vendo a partir da análise daquilo que estamos observando (amostra).

Como você pode perceber, a estatística descritiva e a exploratória são similares, pois ambas objetivam explicitar informações presentes nos dados analisados. Porém, diferem em termos de técnicas e aprofundamento.

A exploração pressupõe investigações profundas a serem posteriormente descritas da melhor forma possível. Dessa maneira, podemos dizer que, enquanto você está efetuando seus cálculos em busca de informações escondidas em seus dados, você está "explorando" e, a partir do momento em que você inicia a explicitação dos resultados, você está "descrevendo-os".

Calcular parâmetros como média e as medidas de variabilidade e explicitá-los pode ser considerado uma exploração e uma descrição dos dados coletados.

Para melhor compreendermos os propósitos da estatística exploratória e descritiva, é necessário conhecermos algumas definições básicas:

Atributos: quando os dados estatísticos apresentam um caráter qualitativo, o levantamento e os estudos necessários ao tratamento desses dados são designados genericamente de estatística de atributo.

Variável: é uma condição ou característica das unidades da população. Por exemplo, a idade das pessoas residentes no Brasil, ou a classe social, são variáveis. As variáveis são classificadas em dois tipos: qualitativas e quantitativas.

Variáveis qualitativas ou por atributos: quando os dados são distribuídos em categorias mutuamente exclusivas. Seus valores são expressos por atributos. São exemplos de variáveis qualitativas: sexo, cor da pele, cidade de nascimento, tipo sanguíneo (O, A, B, AB) etc. Essas variáveis são classificadas em dois tipos:

- **Nominal** *(exemplo: gênero — masculino e feminino);*
- **Ordinal** *(exemplo: classe social — A, B, C, D e E).*

Variáveis quantitativas ou numéricas: Quando os dados são de caráter nitidamente quantitativo e o conjunto dos resultados possui uma estrutura numérica. São exemplos de variáveis quantitativas: idade, estatura, taxa de colesterol etc. As variáveis quantitativas ou numéricas são classificadas em dois tipos:

- **Variável discreta**: a variável discreta só pode assumir apenas valores inteiros. São exemplos de variáveis discretas: número de filhos (0, 1, 2, 3 etc.), número de estudantes em uma sala de aula etc.;
- **Variável contínua**: a variável contínua pode assumir qualquer valor em um dado intervalo. Exemplo de variável contínua: peso de uma pessoa (60,50 Kg).

Dados: São os valores da variável em estudo, obtidos por meio de uma amostragem. Os dados são do mesmo tipo que as variáveis. Por exemplo, uma variável discreta produz dados discretos.

Dados brutos: é uma tabela ou relação de elementos que não foram numericamente organizados. É difícil formarmos uma ideia exata do comportamento do grupo como um todo a partir de dados não ordenados ou dados brutos.

Exemplo: 45, 41, 42, 41, 42 43, 44, 41, 50, 46, 50, 46, 60, 54, 52, 58, 57, 58, 60, 51

ROL: é a tabela obtida após a ordenação dos dados brutos (de forma crescente ou decrescente).

Exemplo: 41, 41, 41, 42, 42 43, 44, 45, 46, 46, 50, 50, 51, 52, 54, 57, 58, 58, 60, 60

A Estatística é dividida basicamente em duas áreas: **estatística descritiva e inferencial**. No nosso curso, serão abrangidas ambas as áreas, conforme veremos.

Estatística descritiva:

Os objetivos da estatística descritiva envolvem coleta, organização e descrição de um conjunto de dados quantitativos ou qualitativos. Com a construção de gráficos, tabelas e com o cálculo de medidas com base em uma coleção de dados numéricos, poderemos compreender melhor o comportamento da variável expressa no conjunto de dados sob análise.

Estatística inferencial:

É a área da estatística responsável pela análise e interpretação dos dados, associada a uma margem de incerteza. Nessa fase, são empregados métodos que tornam possível a estimação de características de uma população baseadas nos resultados amostrais.

A condução de uma pesquisa eleitoral ilustra o processo da inferência estatística. O pesquisador, impossibilitado de entrevistar todos os eleitores (população), seleciona uma amostra de eleitores e questiona sobre suas preferências eleitorais. Baseado nas respostas amostrais, conclui sobre todo o conjunto dos eleitores. Junto com suas conclusões, o pesquisador informa o nível de confiança de que seus resultados amostrais reflitam corretamente o comportamento de todos os eleitores (população).

1.1 Medidas de posição

Como o próprio título sugere, nosso objetivo aqui é a determinação de medidas que ofereçam o posicionamento da distribuição dos valores de uma variável que desejamos analisar.

As medidas de posição são os cálculos estatísticos que representam uma série de dados orientando-nos quanto à posição da distribuição de dados, sendo as medidas de posição mais utilizadas a média aritmética, a moda e a mediana.

Média aritmética ($\bar{x}$)

A medida de tendência central mais comum para um conjunto de dados é a **média aritmética**. A média aritmética amostral de um conjunto de dados é o quociente entre a soma dos valores do conjunto e o número total dos valores.

$$\bar{x} = \frac{\textit{soma dos valores de } x}{\textit{número de observações}} = \frac{\sum x_i}{n}$$

Onde: ***xi*** *são os valores da variável e* ***n*** *o número de valores.*

Exemplo 1: Encontrar a média aritmética para um conjunto de observações: 5, 1, 6, 2, 4.

Solução: *Temos cinco observações: n=5. Então:*

$$\bar{x} = \frac{\sum x_i}{n} = \frac{5+1+6+2+4}{5} = 3,6$$

Cálculo da média para dados agrupados em intervalos de classe

Quando a amostra é muito grande e os dados são discretos, podem ocorrer valores repetidos. Nesse caso, é razoável organizar os dados em uma tabela de distribuição de frequências e trabalharmos com **dados agrupados**.

Quando os dados estiverem **agrupados** em uma distribuição de frequência, usaremos a média aritmética dos valores de $x_1, x_2, x_3, \ldots x_n$ ponderados pelas respectivas frequências absolutas: $F_1, F_2, F_3, \ldots, F_n$. Assim:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$$

Exemplo 2: "neste caso, convencionamos que todos os valores incluídos em um determinado intervalo de classe coincidem com o seu ponto médio e determinamos a média aritmética ponderada por meio da fórmula anterior, onde x_i é o ponto médio da classe" (AMAZONAS, 2013, p. 16). Dada a seguinte distribuição de frequência na Tabela 2.1:

Tabela 2.1 **Tabela de distribuição de frequência**

Intervalos das classes	Frequência (F_i)	x_i	x_i.Fi.
50 |— 54	4	52	208
54 |— 58	9	56	504
58 |— 62	11	60	660
62 |— 66	8	64	512
66 |— 70	5	68	340
70 |— 74	3	72	216
Total	**40**		**2440**

x_i: Ponto médio das classes

Fonte: Amazonas (2013).

Aplicando a equação anterior, temos que:

$$\overline{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{2440}{40} = 61$$

Média Aritmética Ponderada ($\overline{x}$):

A média aritmética ponderada também é chamada de média ponderada. É empregada quando as variáveis têm diferentes importâncias relativas ou, ainda, diferentes pesos relativos.

No cálculo da média ponderada, cada valor coletado na série tem uma participação proporcional ao seu peso, isto é, proporcional à importância relativa no conjunto.

O cálculo da média ponderada é obtido pela soma das variáveis multiplicadas pelos seus pesos, dividida pela soma dos pesos de cada variável. Assim:

$$\overline{x} = \frac{p_1 * x_1 + p_2 * x_2 + p_3 * x_3 + ...p_n * x_n}{p_1 + p_2 + p_3 + ...p_n} = \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i}$$

Onde:

$\overline{x}$ — média ponderada

x_i — observações ou números da variável em estudo

p_i — ponderações ou pesos da variável

Exemplo: Calcular a **média aritmética ponderada** dos números 10, 14, 18 e 30, sabendo-se que os seus pesos são, respectivamente, 1, 2, 3 e 5.

Resolução:

$$\overline{x} = \frac{1*10 + 2*14 + 3*18 + 5*30}{1+2+3+5} = \frac{242}{11} = 22$$

1.1.1 Mediana (Md)

A mediana é o valor que ocupa a posição central do conjunto de dados ordenados (ROL), tal que 50% dos valores são menores que a mediana e os demais 50% são maiores.

Para a sua determinação, utiliza-se a seguinte regra depois de ordenada a amostra de n elementos dispostos segundo uma ordem (crescente ou decrescente): “Quando o número de elementos (n) da série estatística for ímpar, haverá coincidência da mediana com um dos elementos da série” (AMAZONAS, 2013, p. 19). Nesse caso, existirá um único valor de posição central — esse valor será a mediana.

Por exemplo, considere o conjunto de dados {2, 5, 6, 9, 10, 13, 15}. O valor que divide essa série em duas partes iguais é igual a 9. Logo, a mediana é 9.

Quando o número de elementos da série estatística for par, nunca haverá coincidência da mediana com um dos elementos da série de dados. A mediana será sempre a média aritmética dos 2 elementos centrais da série de dados. Por exemplo, no conjunto de dados { 0, 0, 1, 1, 2, 3, 3, 4, 5, 6 }, a mediana será a média aritmética do 5º e do 6º termos da série. Portanto, a mediana será = (2+3) / 2, ou seja, m = 2,50.

96928-978-85-68075-12-8 METODOS_QUATITATIVOS_ALTA_25_06_2014.pdf, page 41 @ Preflight - 14:16:28 - June 30, 2014 - PG-64

1.1.2 Cálculo da mediana em dados agrupados em intervalos de classe (variáveis contínuas)

Para se calcular a mediana em dados agrupados, devemos seguir os seguintes passos:

1º Passo: determinamos as frequências acumuladas (**Σ fi = n**);

2º Passo: calculamos $\frac{n}{2}$; como a variável é contínua, não se preocupe se n é par ou ímpar;

3º Passo: marcamos a classe correspondente à frequência acumulada imediatamente superior à Σ fi / 2. Tal classe será a classe mediana (**classe Md**);

4º Passo: Calculamos a mediana pela seguinte fórmula:

$$Mediana = l_{Md} + \frac{\left[\left(\frac{n}{2} - FAA\right) * h\right]}{F_{Md}}$$

Onde:

l_{Md} = limite inferior da classe mediana.

n = tamanho da amostra ou número de elementos.

FAA= **frequência acumulada** da classe anterior à **classe mediana**.

***h* = amplitude** do intervalo da **classe mediana**.

F_{Md} = **frequência** da **classe mediana.**

Tabela 2.2 **Exemplo — cálculo da mediana para dados agrupados**

Intervalos das classes	Frequência (Fi)	$F_{acumulada}$
35 |— 45	5	5
45 |— 55	12	17
55 |— 65	18	35 (classe Md)
65 |— 75	14	49
75 |— 85	6	55
85 |— 95	3	58
Total	**58**	

Fonte: Do autor (2014).

1° Passo: Calcula-se $\frac{n}{2}$. Como n = 58, temos que $\frac{58}{2} = 29° elemento$.

2° Passo: Identifica-se a classe mediana (Md) pela frequência acumulada. Nesse caso, a classe mediana é a 3°.

3° Passo: Aplica-se a fórmula:

$$Mediana = l_{Md} + \frac{\left[\left(\frac{n}{2} - FAA\right) * h\right]}{F_{Md}}$$

Nesse caso**:** $\boldsymbol{l_{Md}}$ = 55; **n** = 58; **FAA**= 17; **h** = 10, $\boldsymbol{F_{Md}}$ = 18; Logo:

$$Mediana = 55 + \frac{\left[\left(\frac{58}{2} - 17\right) * 10\right]}{18} = 61{,}67$$

1.1.3 Moda (Mo)

Dentre as principais medidas de posição, destaca-se a moda. É o valor da amostra que mais se repete, ou seja, valor que ocorre com maior frequência.

A moda é calculada quando os dados não estão agrupados. Ela é facilmente reconhecida: basta, de acordo com a definição, procurar o valor que mais se repete (AMAZONAS, 2013, p. 17). Por exemplo, no conjunto de dados {7, 8, 9, 10, 10, 10, 11, 12}, a moda é igual a 10.

Há séries que não possuem valor modal, isto é, nas quais nenhum valor aparece mais vezes que outros. Por exemplo, o conjunto de dados {3, 5, 8, 10, 12} não apresenta moda. A série é amodal.

Em outros casos, pode haver dois ou mais valores de concentração. Dizemos, então, que a série tem dois ou mais valores modais. Por exemplo, o conjunto de dados {2, 3, 4, 4, 4, 5, 6, 7, 7, 7, 8, 9} apresenta duas modas: 4 e 7 (AMAZONAS, 2013, p. 17). Nesse caso, a série é bimodal.

Distribuições simples: quando uma tabela de distribuição de frequência apresenta grande quantidade de dados. É importante destacar a classe de maior frequência, a chamada **classe modal**. Essa classe mostra a área em que os dados estão concentrados. Assim, para a distribuição:

Tabela 2.3 **Exemplo — cálculo da moda para dados agrupados**

xi	243	245	248	251	307
Fi	7	17	23	20	8

Fonte: Dos autores (2014).

A moda será 248 (maior frequência), que será indicada por **Mo=248 (moda)**

1.1.4 Cálculo da moda em valores agrupados em intervalos de classe:

A classe que apresenta a maior frequência é denominada **classe modal.** Pela definição, podemos afirmar que a **moda**, nesse caso, é o valor dominante que está **compreendido entre os limites da classe modal**. Um dos métodos para determinação da moda é a aplicação da **fórmula de *CZUBER***:

1° Passo: Identifique a **classe modal** (aquela que possui maior frequência);

2° Passo: Aplicar a fórmula:

$$Mo = l + \left(\frac{d_1}{d_1 + d_2} \right) * h$$

Onde:

l = limite inferior da classe modal

d_1 = frequência da classe modal — frequência da classe anterior à da classe modal

d_2 = frequência da classe modal — frequência da classe posterior à da classe modal

h = amplitude da classe modal

Tabela 2.4 **Exemplo — Cálculo da moda para dados agrupados em distribuição de frequência**

Intervalos das classes	Frequência (Fi)
0 |— 1	3
1 |— 2	10
2 |— 3	17
3 |— 4	8
4 |— 5	5
Total	**43**

Fonte: Dos autores (2014).

1° Passo: Identifica-se a **classe modal**. Nesse caso, trata-se da 3° classe **2 |— 3.**

2° Passo: Sabendo que: ***l* =** *2;* d_1 *= 17-10=7;* d_2 *= 17-8=9;* ***h*** *= 1, aplica-se a fórmula dada acima:*

$$Mo = 2 + \left(\frac{7}{7+9}\right) * 1 = 2{,}44 \text{ (moda estimada)}$$

1.2 Separatrizes

Além das medidas de posição citadas anteriormente, há outras que, consideradas individualmente, **não são medidas de tendência central**, mas estão ligadas à mediana, relacionadas à sua característica de separar a série em duas partes que apresentam o mesmo número de valores. Essas medidas são **os quartis, os decis e os percentis** que são, juntamente com a **mediana**, conhecidas pelo nome genérico de **separatrizes**.

1.2.1 Quartis — Q

Denominamos **quartis** os valores de uma série que a **dividem em quatro partes iguais**. Precisamos, portanto, de **3 quartis (Q1, Q2 e Q3)** para dividir a série em quatro partes iguais.

- Q_1= 1º quartil, deixa 25% dos elementos.
- Q_2= 2º quartil, deixa 50% dos elementos. O quartil 2 (Q2) <u>sempre será igual à mediana da série</u>.
- Q_3= 3º quartil, deixa 75% dos elementos.

1.2.2 Quartis em dados não agrupados

O método mais prático é utilizar **o princípio do cálculo da mediana** para os **3 quartis**. Na realidade, **serão calculadas "3 medianas" em uma mesma série**.

Exemplo 1: Calcule os **quartis** da série: { 5, 2, 6, 9, 10, 13, 15 } **(n=impar)**

- O primeiro passo a ser dado é o da ordenação (crescente ou decrescente) dos valores: { 2, 5, 6, 9, 10, 13, 15 };
- O valor que divide a série acima em duas partes iguais é igual a **9**, logo a **mediana = 9, que será = Q2 = 9**;

- Temos agora {2, 5, 6 } e {10, 13, 15 } como sendo os dois grupos de valores iguais proporcionados pela mediana (**quartil 2**). Para o cálculo do **quartil 1 e 3,** basta **calcular as medianas das partes iguais provenientes da verdadeira mediana da série (quartil 2)**.

Logo:

Em { 2, **5**, 6 }, a mediana é = **5**. Ou seja, será o **quartil 1 = Q1 = 5**.

Em {10, **13**, 15 }, a mediana é =**13**. Ou seja, será o **quartil 3 = Q3 = 13**.

Exemplo 2: Calcule os **quartis** da série: {1, 1, 2, 3, 5, **5, 6**, 7, 9, 9, 10, 13} **(n=par)**

- A série já está ordenada, então calcularemos o **quartil 2 = mediana = (5+6)/2 = 5,5**;
- O **quartil 1** será a mediana da série à esquerda da mediana central**:** {1, 1, **2, 3**, 5, 5} **Q1 = (2+3)/2 = 2,5**;
- O **quartil 3** será a mediana da série à direita de Md: {6, 7, **9, 9**, 10, 13 } **Q3 = (9+9)/2 = 9**.

1.2.3 Quartis para dados agrupados em classes

Usamos a mesma técnica do cálculo da mediana, bastando substituir, na fórmula da mediana,

Σ fi / 2 por k **.** Σ **fi / 4**, sendo **k** o número de ordem do quartil.

Assim, temos:

$$Q_1 = I^* + \frac{\left[\left(\frac{\sum f_i}{4} - FAA\right) * h^*\right]}{f^*}$$

$$Q_2 = I^* + \frac{\left[\left(\frac{2 * \sum f_i}{4} - FAA\right) * h^*\right]}{f^*}$$

$$Q_3 = I^* + \frac{\left[\left(\frac{3 * \sum f_i}{4} - FAA\right) * h^*\right]}{f^*}$$

Onde:

I* = ***limite inferior*** *da* ***classe mediana***.

FAA = ***frequência acumulada*** *da classe anterior à* ***classe mediana***.

f* = ***frequência simples*** *da* ***classe mediana.***

h** = *amplitude *do intervalo da* ***classe mediana****.*

Exemplo 3: *Calcule os quartis da Tabela 2.5*:

Tabela 2.5 **Distribuição de frequência — Cálculo dos Quartis**

Classes	Frequência (f^i)	Frequência acumulada
50 |— 54	4	4
54 |— 58	9	13
58 |— 62	11	24
62 |— 66	8	32
66 |— 70	5	37
70 |— 74	3	40
Total	**40**	

Fonte: Dos autores (2014).

O **quartil 2 = mediana**. Logo:

Σ **fi / 2** = 40 / 2 = 20

Logo, a classe mediana será **58 |— 62**

l* = 58   FAA = 13   f* = 11   h* = 4

$$Q_2 = I^* + \frac{\left[\left(\frac{2 * \sum f_i}{4} - FAA\right) * h^*\right]}{f^*}$$

Substituindo esses valores na fórmula, obtemos:

Md = 58 + [(20 - 13) x 4] / 11 = 58 + 28/11 = **60,54 = Q2**

O **quartil 1**: 1*Σfi / 4 = 10 (correspondente à 2º classe)

$$Q_1 = I^* + \frac{\left[\left(\frac{\sum f_i}{4} - FAA\right) * h^*\right]}{f^*}$$

Q_1 = 54 + [(10 - 4) x 4] / 9 = 54 + 2,66 = **56,66 = Q1**

O **quartil 3**: 3*Σfi / 4 = 30 (correspondente à 4º classe)

$$Q_3 = I^* + \frac{\left[\left(\frac{3*\sum f_i}{4} - FAA\right) * h^*\right]}{f^*}$$

Q_3 = 62 + [(30 -24) x 4] / 8 = 62 + 3 = **65 = Q3**

1.2.4 Decis — D

A definição dos **decis** obedece ao mesmo princípio dos **quartis**, com a modificação da porcentagem de valores que ficam aquém e além do **decil** que se pretende calcular. A fórmula básica será: **k*Σ fi / 10,** onde **k** é o número de ordem do **decil** a ser calculado. Indicamos os **decis: D1, D2,..., D9**. Desse modo, precisamos de **9 decis** para **dividirmos uma série em 10 partes iguais**.

De especial interesse é o **quinto decil,** que ***divide o conjunto em duas partes iguais***. Assim sendo, o **quinto decil é igual ao segundo quartil**, que por sua vez **é igual à mediana**.

Para **D5,** temos: **5*Σ fi / 10 = Σ fi / 2**

Exemplo: Calcule o **3º decil** da tabela anterior com classes.

k = 3, *onde 3*.**Σ** ***fi / 10 =*** *3 x 40 / 10 =* ***12****.*

Esse resultado corresponde à 2ª classe.

D3 *= 54 + [(12 - 4) x 4] / 9 = 54 + 3,55 =* ***57,55 = D3***

1.2.5 Percentil ou centil

Denomina-se **percentis ou centis** os noventa e nove valores que separam uma série em 100 partes iguais. Indicamos: P1, P2, ..., P99. É evidente que **P50 = mediana**; **P25 = Q1** e **P75 = Q3.**

- O cálculo de um **centil** segue a **mesma técnica do cálculo da mediana**, porém a fórmula será **k = Σ fi / 100**, onde **k** é o número de ordem do **centil** a ser calculado.

Questões para reflexão

Considerando o que foi apresentado até o momento nesta unidade, você consegue visualizar uma aplicação dessas ferramentas para medidas de posição no cotidiano do gestor de uma organização?

1.3 Medidas de dispersão

Devido à variabilidade das medidas de tendência central, ainda que consideradas como números que têm a finalidade de representar uma série de dados, elas não podem por si mesmas destacar o grau de homogeneidade ou heterogeneidade que existe entre os valores que compõem o conjunto e, portanto, não bastam para descrever um conjunto de dados. As medidas de tendência central são tanto mais descritivas de um conjunto de dados quanto menor for a variabilidade. Então, quando apresentamos medidas de tendência central para descrever um conjunto de dados, devemos indicar também uma medida de **variabilidade ou dispersão.**

Medidas de dispersão absoluta — Amplitude total (A)

É a diferença entre o maior e o menor valor observados no conjunto de dados, conforme segue:

$$A = Maior\ Valor - Menor\ Valor$$

Exemplo:

Para a série: **10**,12,20,22,25,33,**38**, a amplitude total é dada por:

$$A\,mplitude = 38 - 10 = 28$$

Variância populacional (σ^2) e amostral (S^2)

"A variância é uma medida que tem pouca utilidade como estatística descritiva, porém é extremamente importante na inferência estatística e em combinações de amostras" (AMAZONAS, 2013, p. 24). A definição de variância populacional (s2) é dada por:

$$\sigma^2 = \frac{\sum (x_i - \overline{X})^2 . F_i}{N}$$

Onde:

σ^2 *indica variância populacional e lê-se "sigma" ao quadrado.*

$\bar{x}$ = *média.*

Fi = frequência.

N = tamanho da população.

Para o caso do cálculo da variância amostral (s^2), é conveniente o uso da seguinte fórmula:

$$s^2 = \frac{\sum (x_i - \bar{x})^2 . F_i}{n-1}$$

Onde: $\bar{x}$ = média amostral, n = tamanho da amostra.

Desvio-padrão populacional (σ) e amostral (s)

O desvio-padrão é a "medida de dispersão mais geralmente empregada, pois leva em consideração a totalidade dos valores da variável em estudo. É um indicador de variabilidade bastante estável" (AMAZONAS, 2013, p. 22).

Observando-se a fórmula original para o cálculo da variância, nota-se que ela é uma soma de quadrado. Dessa forma, se a unidade da variável for, por exemplo, metro (m), teremos como resultado metro ao quadrado (m^2). Para se ter a unidade original, necessita-se definir outra medida de dispersão, que é a raiz quadrada da variância — o desvio-padrão. Assim, temos:

$$\sigma = \sqrt{^{2}} \left(\textit{desvio-padrão populacional}\right)$$

$$s \quad \sqrt{s^2} \left(\textit{desvio-padrão amostral}\right)$$

Exemplo 3: calcular a variância amostral (s^2) e o desvio padrão amostral (*s*) da seguinte distribuição amostral

Tabela 2.6 **Exemplo 3**

x_i	5	7	8	9	11
F_i	2	3	5	4	2

Resolução do exemplo 3:

Primeiramente, precisamos do valor da média, conforme vimos anteriormente na *web aula* 3. Então, temos que:

Tabela 2.7 **Exemplo de cálculo da média (Exemplo 3)**

x_i	Frequência (Fi)	$X_i.F_i$
5	2	10
7	3	21
8	5	40
9	4	36
11	2	22
Total	**16**	**129**

Fonte: Dos autores (2014).

Portanto, a média é:

$$\bar{x} = \frac{129}{16} = 8{,}06$$

2º) Cálculo da variância amostral (s^2)

Para calcularmos a variância amostral (s^2), é preciso encontrar o valor de $\Sigma \mathbf{d_i^2 F_i}$. Para tanto, uma nova coluna deverá ser considerada na tabela anterior:

Tabela 2.8 **Cálculo de variância amostral (Exemplo 3)**

x_i	Fi	$X_i.F_i$	$\lvert x_i - \bar{x} \rvert = \lvert d_i \rvert$	$\lvert d_i \rvert^2 .F_i$
5	2	10	$\lvert 5-8{,}06 \rvert = \lvert 3{,}06 \rvert$	18,73
7	3	21	$\lvert 7-8{,}06 \rvert = \lvert 1{,}06 \rvert$	3,37
8	5	40	$\lvert 8-8{,}06 \rvert = \lvert 0{,}06 \rvert$	0,02
9	4	36	$\lvert 9-8{,}06 \rvert = \lvert 0{,}94 \rvert$	3,53
11	2	22	$\lvert 11-8{,}06 \rvert = \lvert 2{,}94 \rvert$	17,29
Total	**16**	**129**		**42,94**

Fonte: Dos autores (2014).

$$s^2 = \frac{\sum (x_i - \bar{x})^2 .F_i}{n-1} = \frac{42{,}94}{16-1} = 2{,}86$$

Portanto,

3º) Cálculo desvio-padrão amostral (s)

$$s = \sqrt{s^2} \text{ , logo: } s = \sqrt{2{,}86} = 1{,}69$$

Resumindo: a distribuição possui média **8,06**. Isto é, seus valores estão em torno de 8,06 e seu grau de dispersão é de **1,69**, medido pelo desvio-padrão.

Medida de dispersão relativa: Coeficiente de variação (*CV*)

Trata-se de uma medida relativa de dispersão útil para comparação em termos relativos do grau de dispersão em torno da média de séries distintas. É dado por:

$$CV = \frac{s}{\overline{x}} * 100$$

Onde: *s = desvio-padrão amostral* e $\overline{x}$ = *média*.

Exemplo 4: numa empresa, o salário médio dos homens é de R$ 4000,00, com desvio-padrão de R$ 1500,00, e o das mulheres é de R$ 3000,00, com desvio-padrão de R$ 1200,00.

Resolução do exemplo 4:

$$CV_{homens} = \frac{s}{\overline{x}} = \frac{1500}{4000} * 100 = 37{,}50\%$$

$$CV_{mulheres} = \frac{s}{\overline{x}} = \frac{1200}{3000} * 100 = 40{,}00\%$$

Logo, podemos concluir que, nessa empresa, os salários das mulheres apresentam maior dispersão relativa que o salário dos homens.

Para saber mais

Você deve estudar mais sobre estatística descritiva lendo o seguinte livro:

GARCIA, R. **Estatística**. São Paulo: Pearson Education do Brasil, 2009.

Leia as páginas 61 e 62, unidade 3.

Atividades de aprendizagem

1. Durante sete meses consecutivos, os rendimentos de certa aplicação financeira foram iguais a 1,20%, 0,90%, 1,50%, 0,70%, 2,10%, 1,40% e 1,80%. Com base nesses dados, calcule o rendimento médio (retorno esperado) e o desvio-padrão amostral (*risco do investimento*).

2. A tabela a seguir apresenta a distribuição das importações de certa empresa. Com base nestes dados, calcular o valor da média das importações dessa empresa.

Tabela 2.9 **Distribuição de frequências — Exercício 2**

Valor importado (R$)	Frequências (f^i)
50.000 |— 60.000	5
60.000 |— 70.000	10
70.000 |— 80.000	20
80.000 |— 90.000	10
90.000 |— 100.000	5
Total	50

Seção 2 Análise exploratória e descritiva

Nesta seção, vamos abordar alguns conceitos relevantes do estudo da análise exploratória de dados conceitos da estatística descritiva. Muitas vezes, ao ouvirmos essas palavras, nos lembramos de índices, taxas e números de um modo geral. Porém, a estatística é muito mais do que isso. Na verdade, a estatística é essencial para nossa vida. Fazemos uso dela todos os dias: quando pesquisamos o preço de uma mercadoria, inferimos de quanto será o aumento do aluguel, conjecturamos o comportamento da população nas urnas na próxima eleição. Seus conceitos já permeiam os mais diversos campos de conhecimento e são indispensáveis no campo científico e no mercado financeiro, por exemplo. Isso ocorre porque a estatística nos permite organizar, descrever, analisar e interpretar dados referentes ao fenômeno que estamos estudando. Dessa forma, ela tem um papel importante no meio acadêmico e profissional, pois nos auxilia na compreensão dos fenômenos que nos cercam. Vamos desvendar essa ciência?

Entende-se por análise exploratória o processo pelo qual aplicamos metodologias quantitativas, técnicas matemáticas e estatísticas com o objetivo de obter informações que possam, porventura, estar ocultas ou "mascaradas" quando fazemos apenas a observação superficial dos dados.

Dentre as técnicas de exploração a regressão que discutiremos neste texto é uma das mais interessantes e úteis.

Quando fazemos a observação superficial de dados, podemos tirar algumas conclusões, como, por exemplo, aumentou, diminuiu, manteve-se etc.; porém, quando fazemos perguntas mais complexas — como "Por que isso ocorreu?", "Em função do que tal fenômeno aconteceu?", "Qual é a tendência futura do comportamento de determinada variável?" etc. —, necessitamos de técnicas mais completas do que a simples observação.

"A análise exploratória de dados nos fornece um extenso repertório de métodos para um estudo detalhado dos dados antes de adaptá-los. Nessa abordagem, a finalidade é obter dos dados a maior quantidade possível de informação que indique modelos plausíveis a serem utilizados numa fase posterior — a análise confirmatória de dados, ou inferência estatística" (MEDRI, 2011, p. 2).

Nas nossas atividades cotidianas, o método é sempre determinante, ou seja, se distribuíssemos três receitas de bolo para três confeiteiras, fornecendo-lhes todos os ingredientes da receita, inclusive com a mesma marca e exatamente o mesmo peso de cada um, certamente, pela lógica, deveriam sair três bolos

iguais. Entretanto, sabe-se que isso não ocorre, e os bolos todos diferem um do outro — inclusive no que tange ao sabor.

Na estatística não é diferente, apesar de existirem as fases do método que apresentaremos a partir de agora. Ao fazer pesquisas, um mesmo método pode diferir, dependendo do pesquisador. Contudo, é importante conhecer, pelo menos, os principais, para que tenhamos ideias de como organizar uma pesquisa.

Portanto, o objetivo deste tópico é que você, acadêmico/a, conheça as cinco fases do método estatístico e aplique-o nas atividades desenvolvidas de agora em diante.

2.1 Organização de dados estatísticos — fases do método estatístico

A estatística descritiva tem por objetivos planejar uma pesquisa, coletar dados, descrever e analisá-los, retirando o maior número possível de informações neles contidos, com o objetivo de utilizá-las nas tomadas de decisão. Seguindo esse raciocínio, a estatística divide o estudo e a análise dos dados em algumas fases, que são descritas a seguir:

1ª fase: definição do problema. Definir objetivos:

Saber exatamente aquilo que se pretende pesquisar é o mesmo que definir corretamente o problema inicial.

2ª fase: planejamento da pesquisa:

Como levantar informações? Que dados deverão ser obtidos? Qual levantamento a ser utilizado? Cronograma de atividades? Custos envolvidos? Seguem abaixo algumas perguntas que precisam ser respondidas no planejamento de um levantamento de dados estatísticos:

- **O quê?** — Características a serem observadas: **variáveis**.
- **Quem?** — Os elementos a serem pesquisados: **população/amostra**.
- **Como?** — O instrumento de coleta de dados: **método de amostragem a ser utilizado, questionário /entrevista estruturada etc.**

3ª fase: execução da pesquisa, coleta de dados:

Fase operacional. É o registro sistemático de dados, focando no objetivo determinado inicialmente.

- **Dados primários (coleta direta):** quando são publicados pela própria pessoa ou organização que os tenha coletado. Exemplos: tabelas do censo demográfico do IBGE, uma empresa que realiza uma pesquisa para saber a preferência dos consumidores pela sua marca.

A **coleta é direta** quando obtida diretamente da fonte. Ex.: empresa que realiza uma pesquisa para saber a preferência dos consumidores pela sua marca.

A coleta direta de dados pode ser classificada em relação ao fator tempo como:

a) **Contínua** (registro) — quando é feita continuamente, tal como a de nascimentos, óbitos, os registros da fiscalização eletrônica de velocidade etc.

b) **Periódica** — quando é feita em intervalos de tempo constantes, como os censos demográficos, matrículas semestrais e/ou anuais dos estudantes etc.

c) **Ocasional** — quando feita ocasionalmente a fim de atender a uma conjuntura ou a uma emergência, como no caso de epidemias, pesquisas eleitorais, avaliações etc.

- **Dados secundários (coleta indireta)**: quando são publicados por outra organização. Exemplo: quando determinado jornal publica estatísticas referentes ao censo demográfico extraídas do IBGE.

Observação: *é mais seguro trabalhar com dados primários. O uso de dados secundários traz o grande risco de erros de transcrição.*

4ª fase: apuração dos dados:

Resumo dos dados através de sua contagem e agrupamento. É a condensação e tabulação de dados.

5ª fase: análise e apresentação dos dados:

Há duas formas de apresentação, que não se excluem mutuamente:

- **Apresentação de dados em tabelas** — é uma apresentação numérica dos dados em linhas e colunas distribuídas de modo ordenado, construídas segundo normas técnicas citadas pela fundação Instituto Brasileiro de Geografia e Estatística (IBGE, 1993).
- **Apresentação gráfica dos dados numéricos** — constitui uma apresentação gráfica, permitindo uma visão rápida e clara da variável estudada.

6ª fase: Interpretação dos dados e conclusões obtidas a partir dos dados (estatística inferencial):

A última fase do trabalho estatístico é a mais importante e delicada. Está ligada essencialmente ao cálculo de medidas e coeficientes, cuja finalidade principal é descrever a variável estudada (estatística descritiva). Na estatística indutiva, ou inferencial, a interpretação dos dados se fundamenta na teoria da

probabilidade, que tem por base a indução, ou inferência, e abstrai, desses resultados, conclusões e previsões.

2.1.1 Método

É o conjunto de meios dispostos convenientemente para chegar a um fim desejado. Dos métodos científicos, vamos destacar o método experimental e o estatístico.

2.1.2 Método experimental

Conforme o próprio termo sugere, esse método consiste em, através da experimentação, manter constantes todas as causas (fatores) menos uma, variando-a de modo que se possa descobrir seus efeitos, caso existam. Esse método é amplamente usado nos diversos campos da atividade humana, bem como nas disciplinas de Física, Química, Biologia etc.

Se você for gerente de uma empresa e permutar um funcionário do setor A com um do setor B e fizer um acompanhamento do resultado, você estará fazendo um método experimental.

2.1.3 Método estatístico

Na impossibilidade de se fazer uso do método experimental, visto que nem sempre temos controle dos fatores envolvidos para fixar seus valores, usamos o método estatístico. O método estatístico considera todas as causas envolvidas no processo como variável e procura determinar, no resultado final, que influências cabem a cada uma delas. Como exemplo, pode-se citar a viabilidade, ou não, do lançamento de determinado produto a partir de uma pesquisa de mercado. Pelo processo do método experimental, nesse caso, poderia ser dispendioso e inadequado (basta imaginar a quantidade de fatores envolvidos nessa pesquisa — não temos como fixar todos eles).

2.1.4 Crítica dos dados

Obtidos os dados, eles devem ser cuidadosamente criticados a fim de procurar-se possíveis falhas e imperfeições, de modo que não se incorra em erros grosseiros que possam influir sensivelmente nos resultados. A crítica dos dados pode ser de origem externa ou interna.

- **Crítica externa:** quando visa atribuir as causas dos erros ao informante, por distração ou má interpretação das perguntas que lhe foram feitas.
- **Crítica interna:** quando visa observar os elementos originais dos dados da coleta.

Questões para reflexão

Ao fazer sua pesquisa, pense num tema que você ache interessante, pois é preciso que você se sinta estimulado para que seus objetivos sejam alcançados. Você já pensou em fazer uma pesquisa? Que tema acha interessante na área de gestão empresarial?

Para saber mais

Você deve estudar mais sobre estatística descritiva lendo o seguinte livro:

DE SOUZA, A. M. S.; GESSER, K.; DALPIAZ, M. V. A. D. Caderno de Estudos — **Estatística**. Editora Uniasselvi, INDAIAL-SC, 2011.

Leia o tópico 2 — páginas 11 a 16.

Atividades de aprendizagem

1. Cite as fases do método estatístico.
2. Para você, o que é coletar dados?
3. Como podem ser apresentados, ou expostos, os dados?

Seção 3 Tabulação e análise de dados em gráficos

Um dos elementos importantes na estatística é o uso de gráficos e tabelas para apresentar os dados estatísticos e seus cálculos. Você pode usar gráficos e tabelas para organizar seus dados e dar mais visibilidade e clareza às informações.

Assim como existem algumas regras e normas que devem ser observadas quando vamos elaborar um texto científico ou acadêmico, para organizar os dados em séries estatísticas, ou em distribuição de frequências, existem algumas normas nacionais definidas pela Associação Brasileira de Normas Técnicas (ABNT) que devem ser respeitadas. Na sua opinião, o uso de gráficos e tabelas para representar os dados estatísticos é importante? Com certeza sim, pois as ferramentas que serão apresentadas nessa seção são essenciais para que vocês organizem de forma correta os dados estatísticos que foram obtidos. Sendo assim, convido todos vocês a explorar mais esta área da estatística tão importante para o futuro gestor. Vamos lá?

3.1 Tabelas

As tabelas podem ser consideradas quadros em que resumimos um conjunto de dados organizados e dispostos sistematicamente em linhas e colunas.

Temos que ter em mente que a tabela deverá ser uma forma objetiva de se demonstrar o comportamento de variáveis; o que se deve buscar são representações simples que possibilitem ao leitor a compreensão do fenômeno sem muito esforço.

Dessa forma, toda tabela estatística deve conter elementos essenciais e elementos complementares, quando necessário. Uma tabela deve apresentar a seguinte estrutura:

3.1.1 Elementos essenciais

- **Título —** no título, devemos indicar a natureza do fato estudado, ou seja, o que foi estudado; também deve conter as variáveis escolhidas na análise do fato, o local e a época em que os dados foram obtidos.
- **Corpo —** é formado pelo conjunto de linhas e colunas em que podemos observar as séries horizontais e verticais de informações.

- **Cabeçalho —** no início de cada coluna, devemos designar a natureza do conteúdo de que a coluna trata.
- **Coluna indicadora —** nessa coluna, devemos evidenciar a natureza do conteúdo de cada linha.

3.1.2 Elementos complementares

- **Fonte —** na fonte, devemos indicar a entidade responsável pela sua organização ou que forneceu os dados primários. Deve ficar no rodapé da tabela.
- **Notas —** quando é necessário algum outro esclarecimento além dos essenciais, eles devem ser colocados em forma de notas no rodapé da tabela.
- **Chamadas** — são colocadas também no rodapé da tabela e são necessárias para esclarecer pormenores ou detalhes em relação às caselas, colunas ou linhas.
- **Atenção:** nenhuma casela deve ficar em branco; elas devem sempre apresentar um número ou sinal. O lado direito e esquerdo das tabelas devem sempre ser abertos.

A Tabela 2.10 apresenta um exemplo de tabela que obedece às normas técnicas.

Tabela 2.10 **População residente no Brasil segundo o sexo, de acordo com o censo demográfico de 2000**

Sexo	População residente
Masculino	83.576.015
Feminino	86.223.155
Total	169.799.170

Fonte: IBGE (2003).

Conforme critério de agrupamento, as tabelas podem representar diversas séries estatísticas, que são descritas a seguir.

a) Série cronológica

É a série estatística em que todos os dados são observados segundo a época de ocorrência. Nessa série, a variável é o tempo, sendo o fato e o local fixos. A Tabela 2.11 apresenta um exemplo de tabela que descreve uma série cronológica.

Tabela 2.11 **Vendas da Companhia Alfa — 2010 a 2014.**

Ano	Vendas (em R$ 1.000,00)
2010	5642
2011	7550
2012	10009
2013	11728
2014	18873

Fonte: Departamento de marketing da Companhia Alfa (dados fictícios).

b) Série geográfica ou de localização

É a série estatística em que os dados são observados segundo a localidade de ocorrência. Nesse tipo de série, a variável é o local e são fixos o fato e a época. Exemplo — Tabela 2.12.

Tabela 2.12 **Vendas de computadores por empresa — 2013**

Empresa controlada	Número de computadores vendidos
A	279
B	321
C	194
D	112
E	228
Total	**1134**

Fonte: Dados fictícios.

c) Série específica

É a série estatística em que os dados são agrupados segundo a modalidade de ocorrência, ou seja, varia o fato e permanece constante a época e o local. Exemplo — Tabela 2.13.

Tabela 2.13 **Regime de trabalho dos colaboradores da empresa Alfa**

Regime de trabalho	Porcentagem
Horista	29%
Tempo integral	35%
Tempo parcial	36%

Fonte: Dados fictícios.

d) Séries conjugadas

Também são conhecidas como tabelas de dupla entrada. Essas tabelas são adequadas quando queremos apresentar duas ou mais séries ao mesmo tempo, de forma que tenhamos duas ordens de classificação: uma horizontal e outra vertical. O exemplo a seguir é de uma série **geográfica-temporal**.

Tabela 2.14 **Vendas relativas ao 1º trimestre de 2013 das filiais da "Roda de Ouro Veículos"**

Filiais	Janeiro/2013	Fevereiro/2013	Março/2013
Paraná	8260	5900	7450
Santa Catarina	6450	6050	5800
Rio Grande do Sul	7300	5600	7200
Total	22000	17550	20450

Fonte: Dados fictícios.

3.2 Apresentação de dados qualitativos

Quando observamos dados qualitativos, classificamos cada unidade da amostra em uma dada categoria. A ideia é resumir as informações na forma de uma tabela que mostre as contagens (frequências) em cada categoria, obtendo, então, uma tabela de distribuição de frequência.

a) Distribuição de frequências

A distribuição de frequências é uma série estatística em que os dados são agrupados com as suas respectivas frequências absolutas.

Nas tabelas de distribuição de frequências, é usual fornecer a proporção (frequência relativa) de unidades que caem em cada categoria. A frequência relativa é dada por:

$$Frequência\ Relativa\ (fr) = \frac{Frequência}{Tamanho\ da\ Amostra}$$

Nas tabelas de distribuição de frequências, é usual fornecer a proporção (frequência relativa) de unidades que pertencem a cada categoria.

A Tabela 2.15 apresenta um exemplo de distribuição de frequência de dados qualitativos.

Tabela 2.15 **Opinião dos consumidores sobre determinado produto**

Respostas	Frequência	Frequência relativa (fr)	Percentual (%)
Bom	1300	$\frac{1300}{2500}=0{,}52$	52%
Regular	450	$\frac{450}{2500}=0{,}18$	18%
Ruim	125	$\frac{125}{2500}=0{,}05$	5%
Não respondeu	625	$\frac{625}{2500}=0{,}25$	25%
Total	2500	1,00	100%

Fonte: Dados fictícios.

3.3 Apresentação de dados quantitativos ou numéricos

Os dados numéricos são apresentados na ordem em que são coletados. Também podem ser apresentados em tabelas de distribuição de frequências (com ou sem intervalos de classe), conforme veremos a seguir.

Distribuição de frequência sem *intervalos de classe*: é a simples condensação dos dados conforme as repetições de seus valores. Para uma amostra de tamanho razoável. Essa distribuição de frequência é inconveniente, já que exige muito espaço.

A Tabela 2.16 apresenta um exemplo de distribuição de frequência de dados quantitativos sem intervalo de classes.

Tabela 2.16 **Distribuição do número de faltas de 30 empregados de uma determinada empresa no semestre**

Número de faltas	Frequência	Percentual (%)
0	9	30,00 %
1	10	33,30 %
2	5	16,70 %
3	3	10,00 %
4	2	6,70 %
5	0	0,00 %
6	1	3,30 %
Total	**30**	**100,00 %**

Fonte: Dados fictícios.

Para uma amostra de tamanho razoável, essa distribuição de frequência é inconveniente, já que exige muito espaço. Tabelas com grande número de dados não oferecem ao leitor visão rápida e global do fenômeno. Por essa razão, tanto dados discretos quanto contínuos, desde que em grande número, devem ser apresentados em tabelas de **distribuição de frequência com intervalos de classes**, conforme veremos a seguir.

Distribuição de frequência *com intervalos de classe*: quando o tamanho da amostra é elevado, é mais indicado efetuar o agrupamento dos valores em vários intervalos de classe. Para construir uma tabela de distribuição de frequência com intervalos de classe, siga os procedimentos a seguir:

1º Passo: organize os dados brutos em um **ROL**;

2º Passo: encontre o valor máximo e mínimo do conjunto de dados e calcule a Amplitude total **(A)**, que é a diferença entre os valores máximo e mínimo. Portanto:

$$\mathrm{A} = \textit{Maior Valor} - \textit{Menor Valor}$$

3º Passo: Calcule o número de classes **(K)**: o número de classes de uma representação será um número inteiro próximo de **K,** que pode ser obtido por vários métodos, sendo os mais usuais:

Regra de Sturges:

$$K = 1 + 3{,}3 * \log_{10} n$$

*Onde: **K** é o número de classes e **n** é o número de dados (tamanho da amostra)*.

O indicado é sempre o arredondamento do valor de K obtido para um valor mais alto.

Observação: o cálculo de **K** por meio de fórmulas pode servir como referência, mas não deve ser entendido como obrigatório.

4º Passo: decidido o números de classes, calcule, então, o tamanho do intervalo de classe **(h),** que é definido por:

$$h = \frac{\mathrm{A}}{K}$$

Assim como no caso do número de classes (**K**), o tamanho do intervalo de classe (**h**) deve ser aproximado para o maior valor inteiro. Por exemplo, se K = 6,61, usa-se **K** = **7**; e se h = 1,7, usa-se **h = 2**.

5º Passo: organize as classes de maneira que a primeira contenha o menor valor observado. O primeiro elemento das classes seguintes sempre será formado pelo último elemento da classe anterior.

Exemplo: vamos construir uma tabela de distribuição de frequências das idades dos funcionários de uma amostra de 50 elementos selecionados de uma empresa.

Dados brutos:

62	56	20	48	22	54	25	65	37	27	37
29	30	30	45	31	54	44	34	35	44	36
37	26	37	29	38	38	38	40	41	43	33
36	45	32	31	46	51	21	58	50	47	53
24	45	20	49	18	25					

1º Passo: a partir dos dados brutos, construir o **ROL** (ordenação dos dados em ordem crescente):

18	20	20	21	22	24	25	25	26	27	29
29	30	30	31	31	32	33	34	35	36	36
37	37	37	37	38	38	38	40	41	43	44
44	45	45	45	46	47	48	49	50	51	53
54	54	56	58	62	65					

2º Passo: determinar a Amplitude total **(A)**:

$$\text{A} = \text{Maior Valor} - \text{Menor Valor}$$
$$\text{A} = 65\text{-}18 = 47$$

3º Passo: Como os dados serão agrupados em classes, é preciso escolher o número de classes **(K)**:

Pela **Regra de Sturges,** temos que: $K = 1 + 3{,}3 * \log_{10} n$, sendo que n=50. Portanto:

$$K = 1 + 3{,}3 * \log_{10}(50) = 1 + 3{,}3 * (1{,}7) = 6{,}61 \cong 7$$

4º Passo: Cálculo do tamanho do intervalo de classe **(h)**:

$$h = \frac{A}{K} = \frac{47}{7} = 6{,}71 \cong 7$$

Quanto aos limites das classes, utilizaremos o seguinte critério **a |— b** (incluiremos nessa classe todos os elementos maiores ou iguais a ***a*** e menores do que ***b***). A Tabela 2.17 mostra um exemplo de uma tabela de distribuição de frequência para variável contínua.

Tabela 2.17 **Distribuição de frequências que representa a idade dos funcionários de certa empresa (n=50)**

Classes	Intervalos das classes	Frequência (Fi)	Frequência relativa (fi)	%
1	18 |— 25	6	6/50= 0,12	12
2	25 |— 32	10	10/50 = 0,20	20
3	32 |— 39	13	13/50 = 0,26	26
4	39 |— 46	8	8/50 = 0,16	16
5	46 |— 53	6	6/50 = 0,12	12
6	53 |— 60	5	5/50 = 0,10	10
7	60 |— 67	2	2/50 = 0,04	4
Soma		**50**	**1**	**100**

Fonte: Dados fictícios.

3.3.1 Apresentação de dados em gráficos

Quando queremos representar visualmente os dados estatísticos, podemos fazer isso utilizando as representações gráficas. Mas, atenção: as representações gráficas devem sempre corresponder a uma tabela, sem, contudo, substituí-la. Os gráficos devem ter as seguintes características: devem obedecer a uma escala, usar o sistema de coordenadas, ter simplicidade, ser claro e, principalmente, deve corresponder à veracidade dos dados.

Gráficos de informação: esses gráficos são usados geralmente quando queremos proporcionar ao público em geral uma visualização rápida e clara. Como são gráficos caracteristicamente expositivos, podemos dispensar comentários adicionais. Também podemos omitir as legendas, desde que as informações relevantes estejam presentes no gráfico.

Gráficos de análise: esses gráficos são mais adequados ao trabalho com a estatística, pois fornecem elementos úteis para a análise dos dados, além de ser também informativo. Normalmente, os gráficos de análise são acompanhados de sua respectiva tabela estatística. Também podemos incluir um texto explicativo, que tem como objetivo esclarecer o leitor dos pontos principais divulgados no gráfico.

O uso de gráficos é muito apreciado em nossa sociedade pelo fato de permitir uma visualização rápida e eficiente. Mas devemos ficar atentos, uma vez que um gráfico mal construído pode transmitir uma informação deturpada em relação à informação verdadeira. Normalmente, isso ocorre por problemas de escala, em que as proporções entre os dados não são respeitadas.

Diagramas: os diagramas são gráficos geométricos que apresentam duas dimensões. São os mais utilizados para representar séries estatísticas.

Vamos ver agora algumas classificações de gráficos:

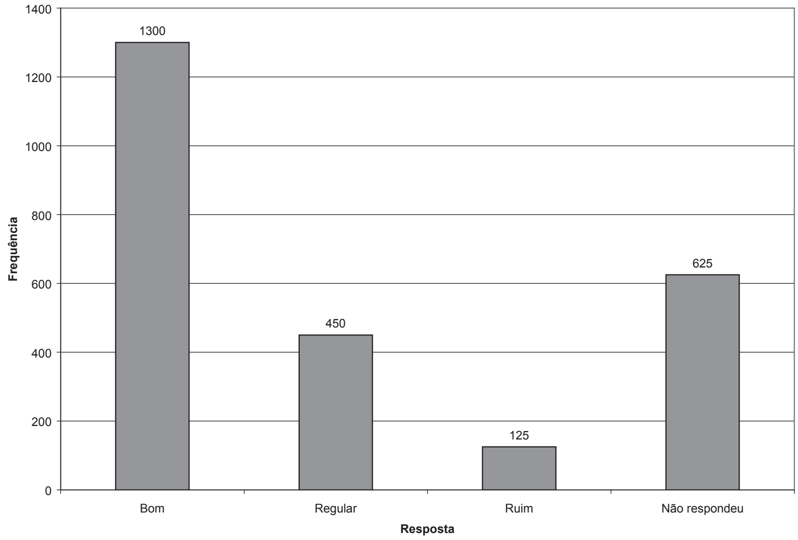
Todo gráfico deve apresentar **título** e **escala**. O **título** deve ser colocado abaixo do gráfico. As **escalas** devem ser crescentes da esquerda para a direita e de baixo para cima.

3.3.2 Gráficos de colunas e barras

Quando as legendas necessárias para a compreensão dos dados não são breves, usamos preferencialmente os gráficos em barras horizontais. Nesse tipo de gráfico, os retângulos têm a mesma base, e as alturas são proporcionais ao valor quantitativo dos dados. A ordem a ser respeitada é a ordem cronológica, no caso de séries temporais, e a decrescente, se a série for geográfica ou categórica. Vamos ver alguns exemplos:

No gráfico de colunas, as barras são apresentadas na posição vertical. Para ilustrar o gráfico de colunas, serão utilizados os dados apresentados na Tabela 2.17, originando, assim, a Figura 2.1 (*gráfico de colunas com linhas auxiliares — grades e rótulos*).

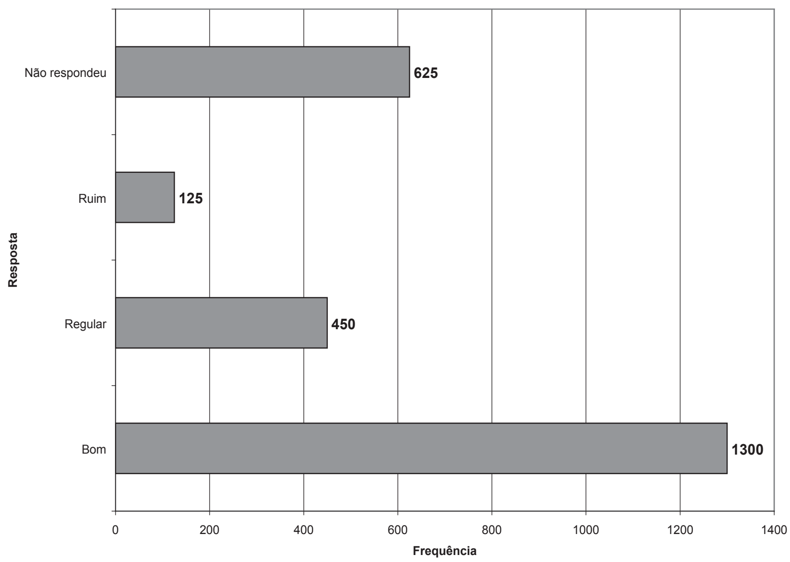
Figura 2.1 **Opinião dos consumidores sobre determinado produto (gráfico de colunas)**



Fonte: Dos autores (2014).

No gráfico de barras, as barras são apresentadas na posição horizontal, como apresentado na Figura 2.2.

Figura 2.2 **Opinião dos consumidores sobre determinado produto (gráfico de barras)**



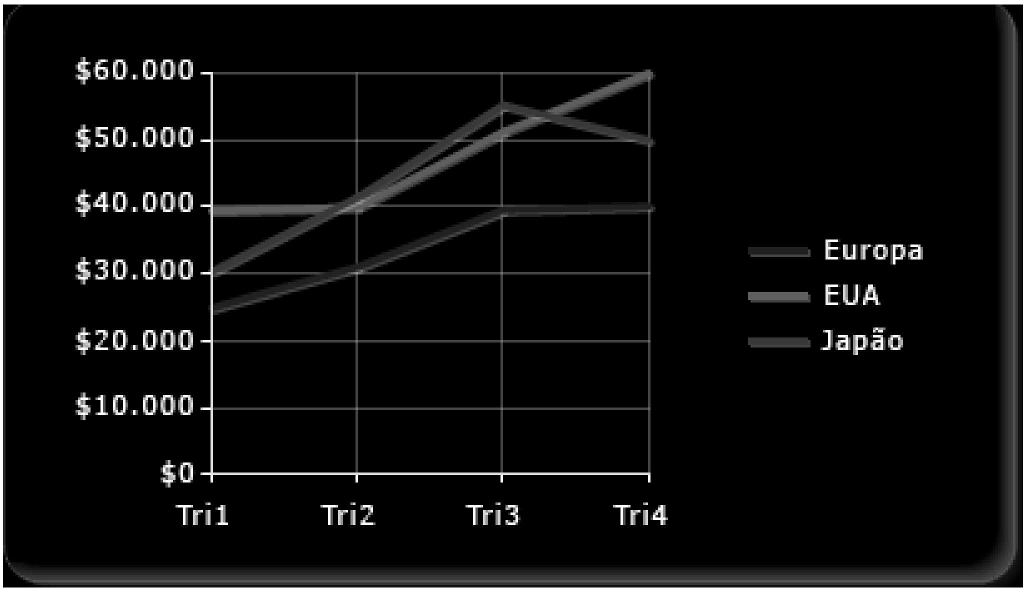
Fonte: Dos autores (2014).

3.3.3 Gráficos de linhas e lispersão

Os gráficos de linhas e dispersão exibem uma série como um conjunto de pontos conectados por uma única linha. As linhas dos gráficos são usadas para representar grandes quantidades de dados que ocorrem em um período de tempo contínuo (séries cronológicas). É o gráfico que melhor representa a evolução conjunta de duas variáveis quantitativas, sendo X considerada a variável independente e Y a variável dependente.

A Figura 2.3 ilustra um gráfico de linhas que contém três séries.

Figura 2.3 **Exemplo do gráfico de linhas**



Fonte: Microsoft (2014).

3.3.4 Gráficos de setores ou pizza

É a representação gráfica de uma série estatística em um círculo, por meio de setores. É utilizado quando se pretende comparar cada valor da série com o total.

A partir dos dados apresentados na Tabela 2.9, foi obtido o gráfico de setores apresentado na Figura 2.4.

Figura 2.4 **Vendas da Companhia Alfa — 2004-2008 (gráficos e setores)**

Fonte: Dos autores (2014).

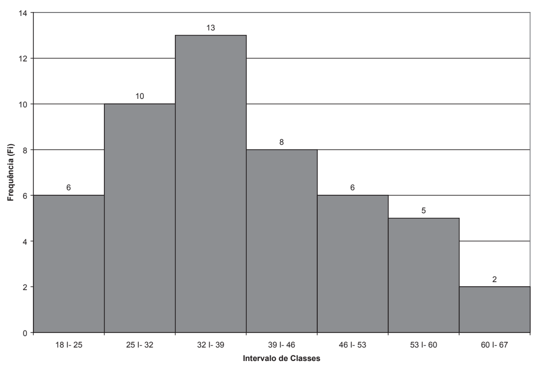
3.3.5 Histograma

É o gráfico que melhor apresenta as frequências de uma variável quantitativa contínua agrupada em classes. Quando os dados são contínuos e a amostra é grande, é mais conveniente condensar os dados, isto é, organizar uma tabela de distribuição de frequências, agrupar os dados em classes e, a partir disso, desenhar um histograma.

O histograma é a representação gráfica de uma distribuição de frequência, formado por um conjunto de retângulos justapostos cujas bases se localizam sobre o eixo horizontal, de tal modo que seus pontos médios coincidam com os pontos médios dos intervalos de classe.

Para ilustrar o histograma, serão utilizados os dados apresentados na Tabela 2.15, obtendo-se, assim, a Figura 2.5.

Figura 2.5 **Histograma para idade de 50 funcionários de certa empresa**



Fonte: Dos autores (2014).

O histograma dispõe de informações de modo que seja possível a visualização da forma de distribuição do conjunto de dados, além da percepção do valor central e da dispersão dos dados em torno desse valor central. **Pelos dados da Figura 2.5, fica fácil perceber que a maior quantidade de funcionários tem idade entre 32 e 38 anos.**

O histograma contém as mesmas informações da tabela de distribuição de frequências. São representações que buscam a organização de grupos de dados quantitativos.

3.3.6 Pictogramas

São os gráficos construídos a partir de figuras que representam a intensidade do fenômeno. Como sua forma é atraente, ele é muito utilizado para chamar a atenção do público em geral, podendo ser muito utilizado em Marketing. Segundo Bruni (2008), alguns aspectos relacionados aos pictogramas devem ser sempre observados:

a) Os símbolos devem ser autoexplicativos;
b) As diferentes quantidades devem expressar-se mediante maior ou menor número de símbolos, e não mediante um aumento ou diminuição do símbolo básico;
c) Os gráficos devem proporcionar uma visão geral do fenômeno, e não detalhes minuciosos;
d) Os pictogramas estabelecem comparações gerais, devendo ser evitados para interpretar afirmações ou dados isolados.

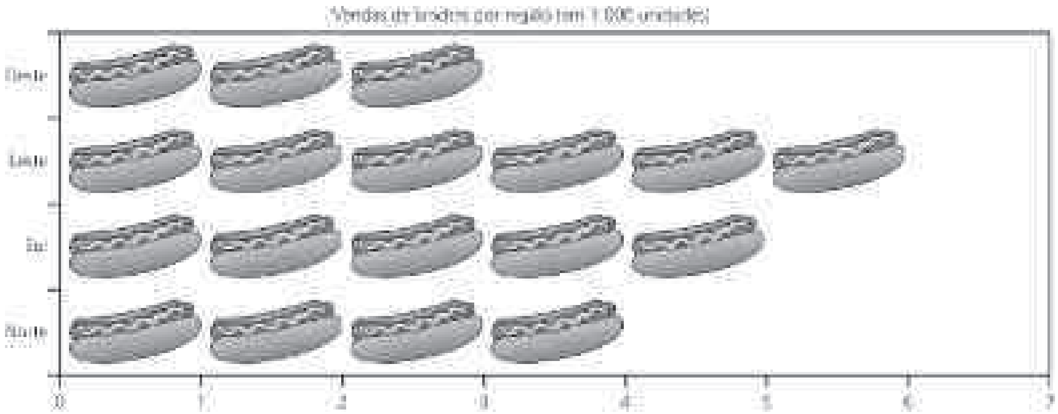
Tabela 2.18 **Venda de lanches por região (em 1000 unidades)**

Região	Vendas
Norte	4
Sul	5
Leste	6
Oeste	3

Fonte: Bruni (2008).

Para ilustrar o pictograma, serão utilizados os dados apresentados na Tabela 2.18, obtendo-se assim a Figura 2.6.

Figura 2.6 **Exemplo de pictograma**



Fonte: Adaptada de Bruni (2008).

1. Na planilha, organize os dados que você deseja plotar em um gráfico.

   Os dados podem ser organizados em linhas ou colunas — o Excel determina automaticamente a melhor maneira de plotá-los no gráfico. Alguns tipos de gráfico (como gráficos de pizza e de bolhas) exigem uma organização específica dos dados.

2. Selecione as células que contêm os dados que você deseja usar no gráfico.

   Dica: se você selecionar apenas uma célula, o Excel plotará automaticamente todas as células que contêm dados adjacentes a essa célula em um gráfico. Se as células que você deseja plotar em um gráfico não estiverem em intervalo contínuo, será possível selecionar células não adjacentes ou intervalos, até que a seleção forme um retângulo. Você também pode ocultar as linhas ou as colunas que não deseja plotar no gráfico.

3. Na guia **Inserir**, no grupo **Gráficos**, siga um destes procedimentos:

- Clique no tipo de gráfico e, em seguida, clique no subtipo de gráfico que deseja usar.
- Para ver todos os tipos de gráfico disponíveis, clique em para iniciar a caixa de diálogo **Inserir Gráfico** (Figura 2.7). Em seguida, clique nas setas para rolar entre os tipos de gráfico.

Figura 2.7 **Inserir gráficos — Software Excel→**

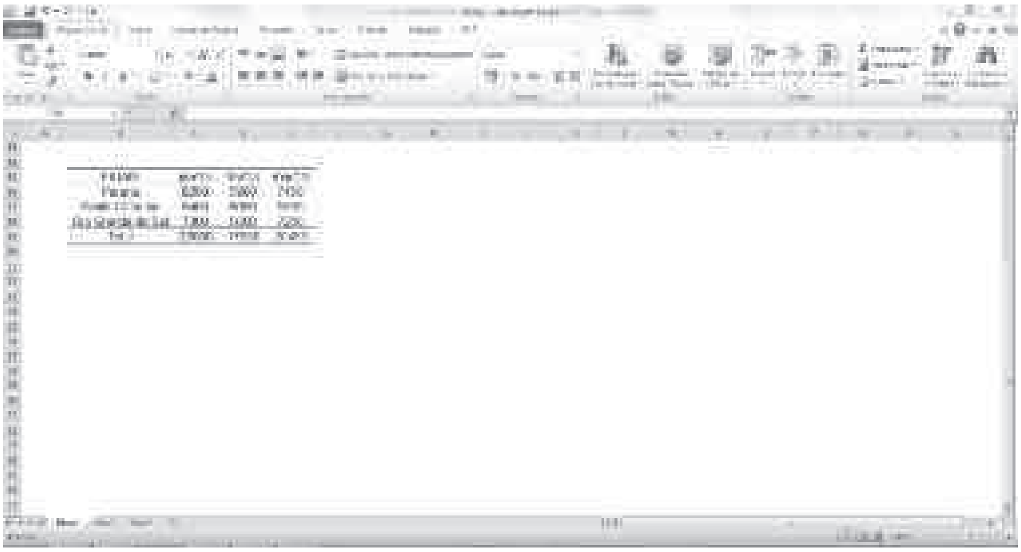
Exemplos:

Na elaboração dos gráficos a seguir, utilizando o software Excel, utilizaremos os dados contidos na Tabela 2.12 dessa unidade.

a) **Exemplo 1 — Gráfico de linhas**

Inserir os dados na planilha do Excel conforme exposto na Figura 2.8:

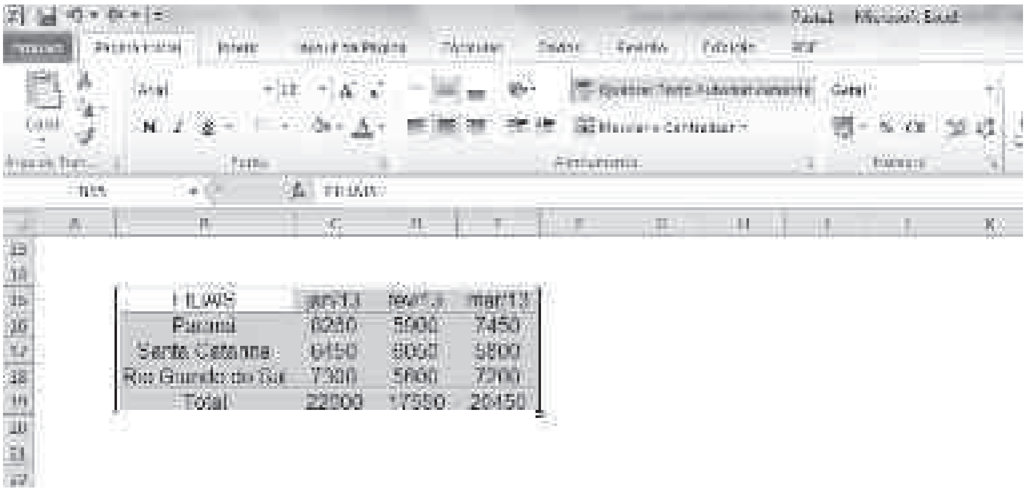
Figura 2.8 **Dados da Tabela 2.14 — Planilha do Excel**



Fonte: Dos autores (2014).

Com o *mouse,* selecionar todos os dados a serem plotados no gráfico, conforme indicado na Figura 2.9.

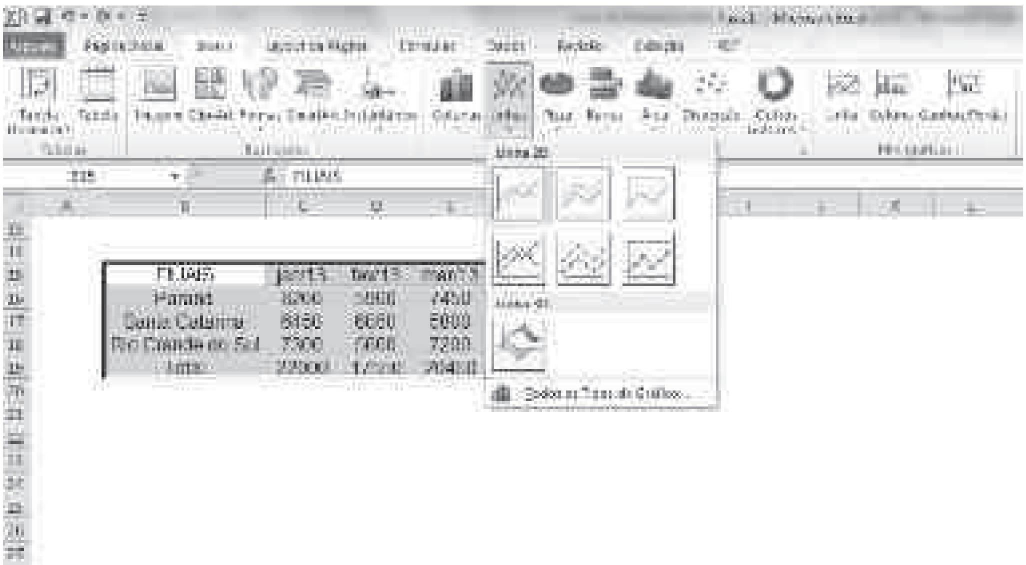
Figura 2.9 **Planilha do Excel — seleção dos dados**



Fonte: Dos autores (2014).

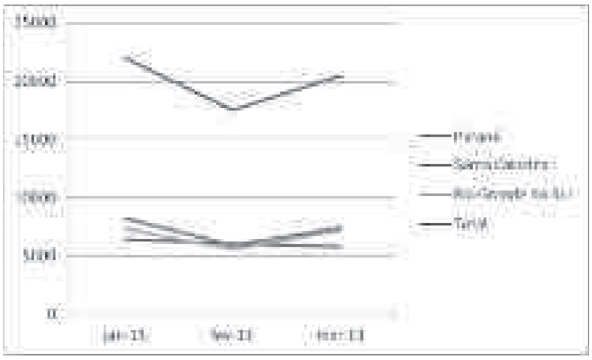
Na guia **Inserir** → **Gráficos,** selecionar a opção **linhas** (Figura 2.10), obtendo--se, assim, o gráfico de linhas apresentado na Figura 2.11.

Figura 2.10 **Planilha do Excel — gráfico de linhas**



Fonte: Dos autores (2014).

Figura 2.11 **Gráfico de linhas elaborado no Excel®**



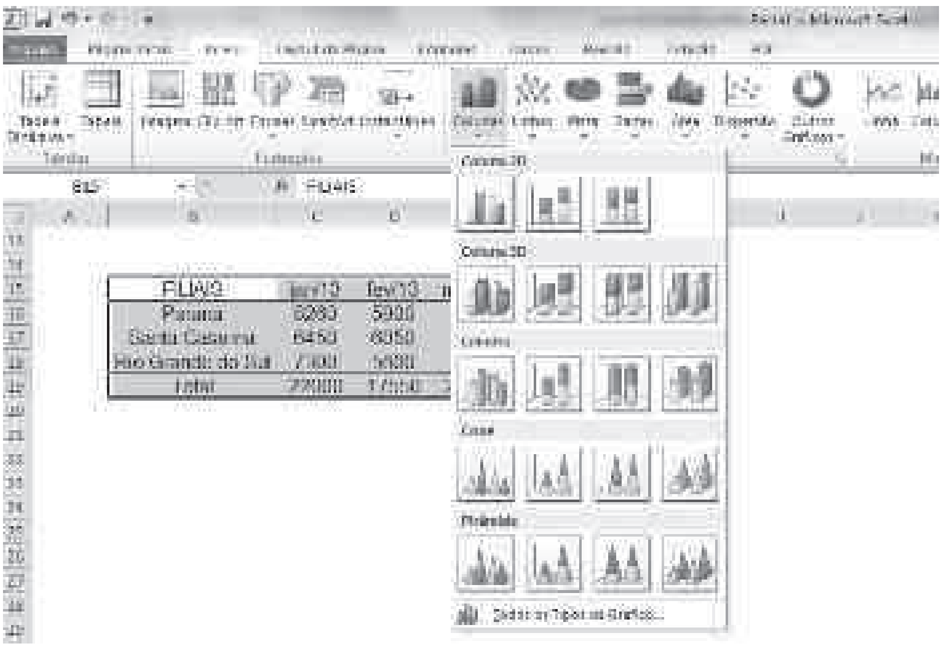
Fonte: Dos autores (2014).

b) Exemplo 2 — gráfico de colunas

A inserção dos dados na planilha do Excel segue o mesmo padrão já utilizado no **exemplo a.**

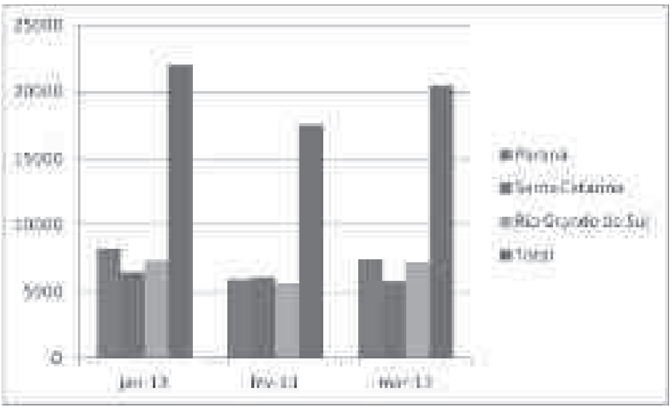
Na guia **Inserir** → **Gráficos,** selecionar a opção **colunas** (Figura 2.12), obtendo-se, assim, o gráfico de linhas apresentado na Figura 2.13.

Figura 2.12 **Planilha do Excel — gráfico de colunas**



Fonte: Dos autores (2014).

Figura 2.13 **Gráfico de colunas elaborado no Excel®**



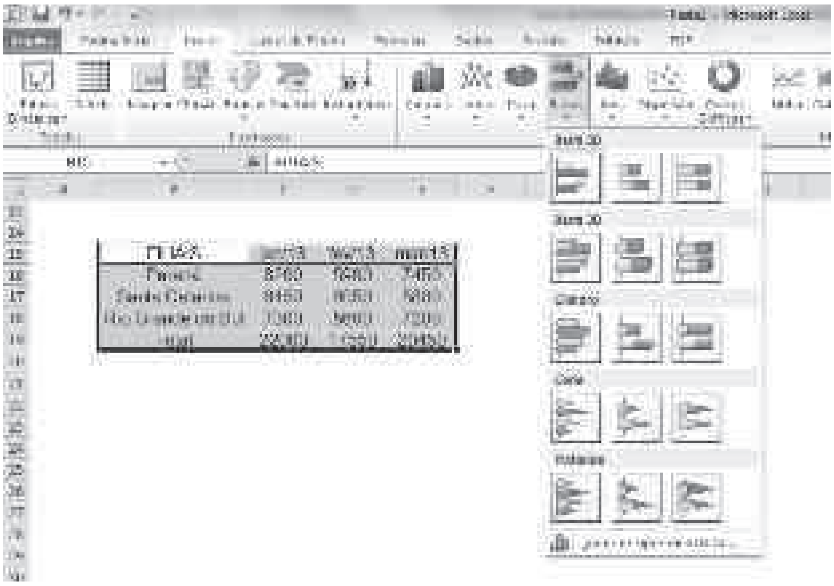
Fonte: Dos autores (2014).

c) Exemplo 3 — gráfico de barras

A inserção dos dados na planilha do Excel segue o mesmo padrão já utilizado no **exemplo a.**

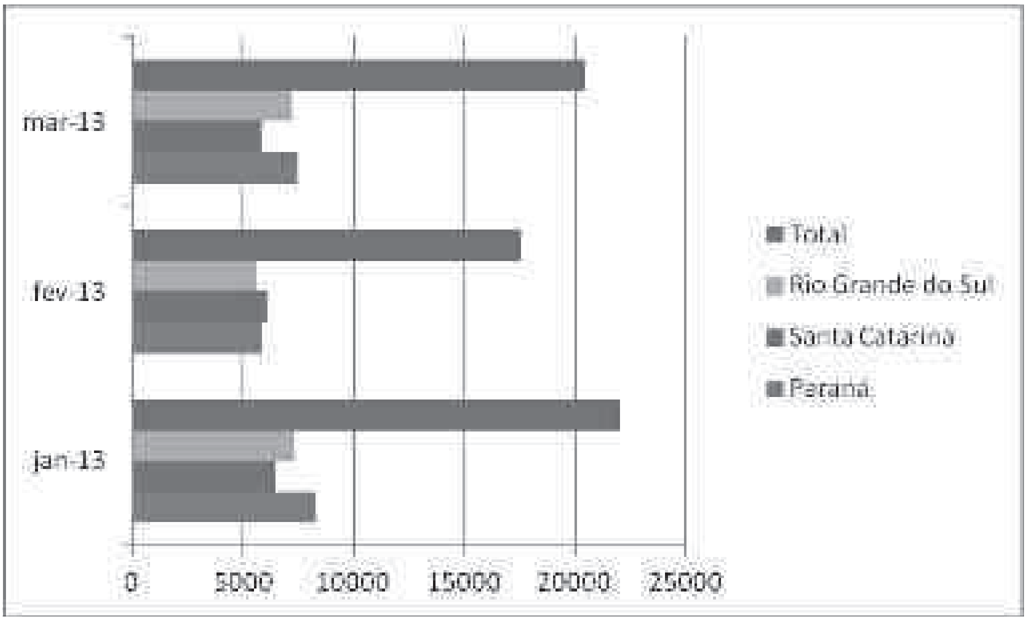
Na guia **Inserir** → **Gráficos,** selecionar a opção **barras** (Figura 2.14), obtendo-se, assim, o gráfico de linhas apresentado na Figura 2.15.

Figura 2.14 **Planilha do Excel — gráfico de barras**



Fonte: Dos autores (2014).

Figura 2.15 **Gráfico de barras elaborado no Excel®**



Fonte: Dos autores (2014).

3.3.7 Diagrama ou gráfico boxplot

Segundo Bruni (2008), o *boxplot*, ou caixa de dados, é um dos mais usuais gráficos da estatística. Representa a dispersão dos dados, revelando a mediana e os quartis — medidas de posição que foram apresentadas na seção 1 desta unidade.

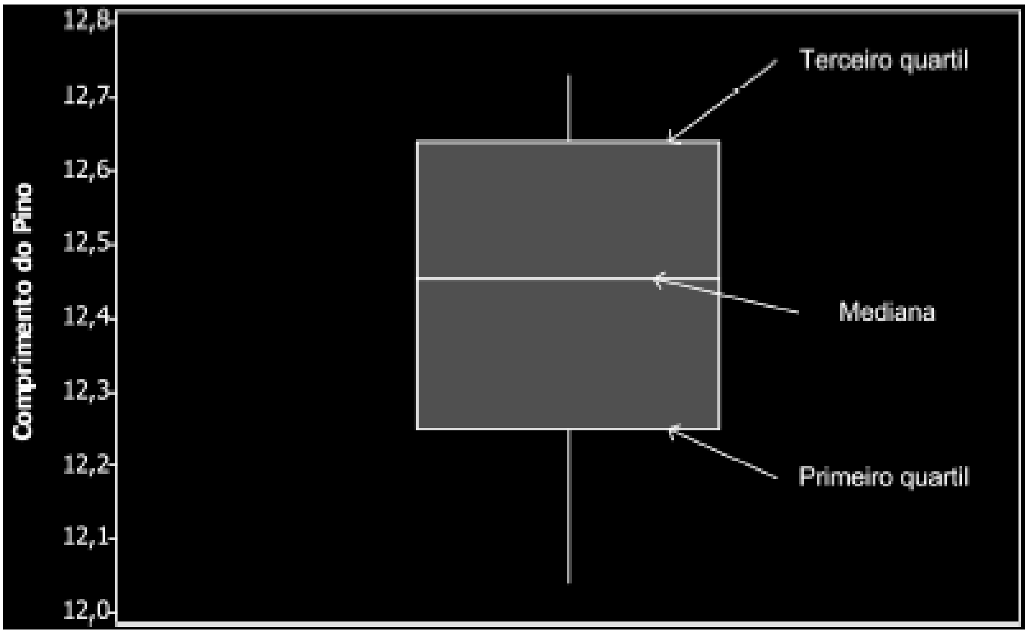
O *boxplot* é um gráfico utilizado para avaliar a distribuição empírica dos dados. Ele é formado pelo primeiro e terceiro quartil e pela mediana. As hastes inferiores e superiores se estendem, respectivamente, do quartil inferior até o menor valor não inferior ao limite inferior, e do quartil superior até o maior valor não superior ao limite superior. Os limites são calculados da forma abaixo:

Limite inferior: $Q_1 - 1{,}5(Q_3 - Q_1)$

Limite superior: $Q_3 - 1{,}5Q_3 - Q_1)$

A Figura 2.16 apresenta um esquema genérico de um *boxplot*.

Figura 2.16 **Exemplo de *boxplot*, separado por grupos**



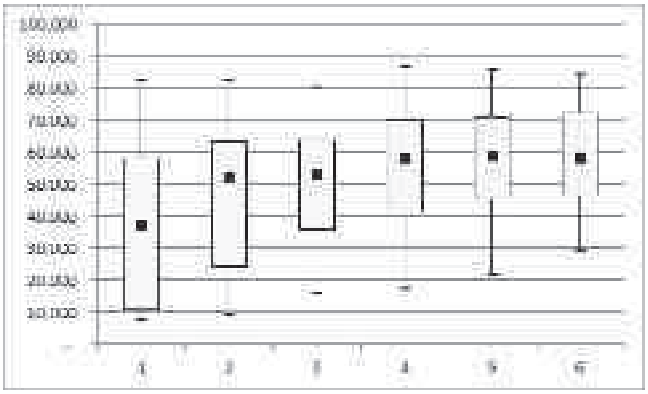
Fonte: Guia... (2012).

Através de uma representação simples, é possível verificar a posição central do conjunto ordenado dos dados, denominado mediana, e as subdivisões das metades das séries ordenadas, denominadas quartis (BRUNI, 2008).

A base do retângulo central é representada pelo primeiro quartil. Abaixo desse ponto, estão situadas 25% das observações na série ordenada. A caixa costuma ser dividida por um segmento de reta, que representa exatamente a mediana — separatriz ou mediana de ordenamento, que deixa 50% das observações da série ordenada abaixo e 50% acima. O topo da caixa correspondente ao terceiro quartil — abaixo deste situam-se 75% das observações e, acima, 25%. Os seguimentos de reta horizontais representam os valores máximos e mínimos da série ordenada (BRUNI, 2008).

A Figura 2.17 ilustra um gráfico de *boxplot* dos pontos obtidos por uma amostra.

Figura 2.17 **Exemplo de *Boxplot***



Fonte: Guia... (2012).

O boxplot pode, ainda, ser utilizado para uma comparação visual entre dois ou mais grupos. Por exemplo, duas ou mais caixas são colocadas lado a lado e se compara a variabilidade entre elas, a mediana e assim por diante. Outro ponto importante é a diferença entre os quartis ($Q_3 - Q_1$), que é uma medida da variabilidade dos dados.

3.3.8 Como criar um gráfico de boxplot utilizando o software Excel®

1. Para criar o gráfico de *boxplot*, crie uma tabela conforme o modelo abaixo e siga o roteiro.

	1	2	3	4	5	6
máximo	82.269	82.275	80.110	86.423	85.654	84.300
Q3	57.840	63.619	63.849	70.048	71.021	72.200
média	37.597	52.645	52.798	58.059	58.800	58.456
mediana	39.873	51.755	51.897	55.717	56.748	53.992
Q1	10.968	24.103	35.792	41.168	45.467	46.346
Mínimo	7.653	9.044	15.721	17.137	21.789	28.956

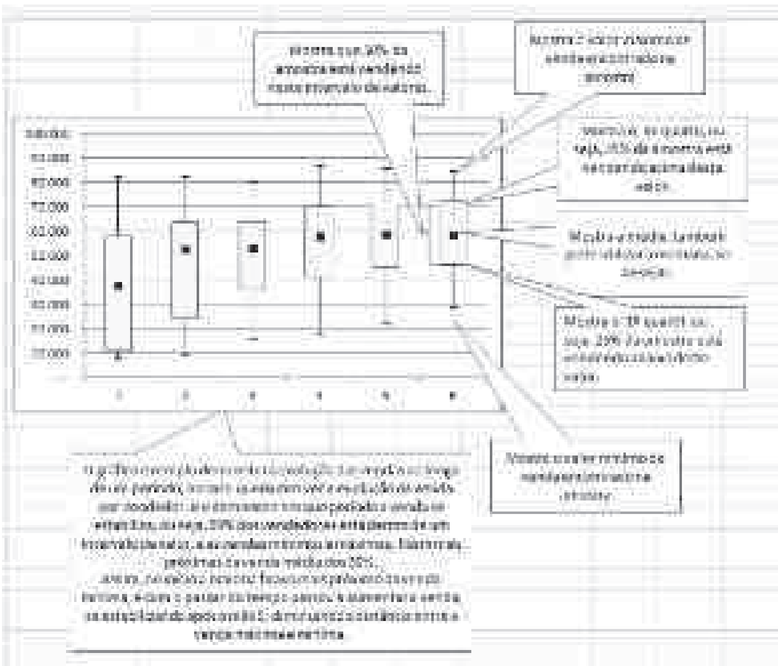
Fonte: Guia... (2012).

2. Crie um gráfico de linhas no Excel, sendo uma linha para cada informação.
3. Na aba **Padrões**, selecione **Linha Nenhuma** e, para **Marcador,** escolha o traço **(—).**

4. Na aba **Opções**, selecione **Linhas de Máximo/Mínimo** e **Barras superiores/inferiores.**
5. Perceba que tem uma ordem nas entradas Q1, Máximo, Média, Mínimo e Q3. Portanto, utilize as setinhas nesta tela do Excel para ajustar a ordem.

6. Na Figura 2.18, a explicação do resultado do gráfico Excel criado a partir do exemplo:

Figura 2.18 Exemplo *boxplot (utilizando o software Excel)*



Fonte: Guia... (2012).

Questões para reflexão

Quando fazemos um gráfico, precisamos nos atentar para que ele apresente uma visualização rápida do fenômeno, que tenha formato adequado à sua publicação em trabalhos técnico-científicos — seja em revistas, periódicos etc. — para que ele realmente forneça as informações apresentadas. Quais gráficos você considerou mais adequados e claros para a apresentação de dados estatísticos? Em revistas e jornais da sua área de atuação, que tipos são mais comuns?

Para saber mais

Você deve estudar mais sobre estatística descritiva lendo o seguinte livro:

BRUNI, A. L. **Estatística aplicada à gestão empresarial**. 2. ed. São Paulo: Atlas, 2008.

Leia a unidade 2 (páginas 23 a 31).

Atividades de aprendizagem

1. A empresa Preço Bom Ltda. registrou o seguinte número de faltas de funcionários em determinado período. A partir destes dados, construa o histograma.

Número de Faltas	F_i
3	2
4	6
5	8
6	5
7	2
Soma	**23**

2. A indústria Qualidade Ltda. verificou que uma amostra formada por 360 funcionários apresentava a seguinte distribuição de escolaridade. Com base nestes dados, construa um diagrama de colunas.

Escolaridade	F_i
Superior	34
Médio	56
Fundamental	112
Técnico	158
Soma	**360**

Fique ligado!

Nesta unidade, você viu que:

- A estatística descritiva pode ser definida como um conjunto de técnicas destinadas a descrever e resumir dados a fim de que possamos tirar conclusões a respeito de características de interesse.
- O método estatístico é composto por cinco fases: a coleta, a crítica, a apuração, a apresentação e a análise dos dados.
- A coleta pode ser direta ou indireta.
- A análise e a interpretação dos dados estatísticos tornam possível o diagnóstico de uma empresa, o conhecimento de seus problemas (condições de funcionamento, produtividade), a formulação de soluções apropriadas e um planejamento objetivo de ação.
- Os gráficos devem obedecer a uma escala, ter simplicidade, ser claro e, principalmente, devem corresponder à veracidade dos dados.
- Uma distribuição de frequência pode ser feita com ou sem intervalos de classes.
- O gráfico indicado para apresentar dados de uma distribuição de frequência com intervalos de classe é o histograma.

Para concluir o estudo da unidade

Caros alunos (as), agora que finalizamos esta unidade, indico que sejam feitos os exercícios de aprendizagem indicados a seguir e também que vocês utilizem os temas abordados nesta unidade em suas atividades diárias, avaliando a sua aplicabilidade nas funções do gestor ou administrador como forma de aprofundar os estudos sobre esses temas.

Atividades de aprendizagem da unidade

1. Dada a tabela de distribuição de frequência a seguir, calcule o valor da mediana.

Intervalos das classes	**Frequência (Fi)**
10 |— 14	12
14 |— 18	25
18 |— 22	36
22 |— 26	35
26 |— 30	22
30 |— 34	18
34 |— 38	8
38 |— 42	4
Total	**160**

2. Considere a tabela de frequências abaixo, que apresenta a distribuição dos salários dos funcionários de certa empresa. Com base nestes dados, calcule o valor da média deste conjunto de dados.

Salários (R$)	**Frequências (f_i)**
1000 |— 2000	25
2000 |— 3000	20
3000 |— 4000	15
4000 |— 5000	13
5000 |— 6000	7
Total	**80**

3. A distribuição abaixo representa o consumo (em Kg) de um determinado produto colocado em oferta em um hipermercado, que limitou o consumo máximo por cliente em 5 kg. Com base nestes dados, calcule a moda (método de Czuber).

Classe	Consumo (em Kg)	N° de clientes
1	0 |— 1	12
2	1 |— 2	15
3	2 |— 3	21
4	3 |— 4	32
5	4 |— 5	54

4. Considerando o seguinte conjunto de dados {5, 4, 8, 10, 12}, calcule o desvio-padrão amostral.

5. Um aluno do curso de Administração está estudando em seu TCC o perfil dos participantes nos eventos da sua área, sendo que uma das variáveis pesquisadas foi a idade das pessoas que participavam dos eventos. No total, foram entrevistadas 45 pessoas, sendo que a idade variou de 18 a 54 anos. Os resultados obtidos foram tabulados em classes de frequências, construídas conforme os procedimentos formais da estatística. Para construção da tabela de distribuição de frequências, a amplitude do intervalo de classes (h) deverá ser:

Obs. Calcular o número de classes **(K)** pela Regra de Sturges ($\log_{10}$ 45=1,65). *Os valores de **h** e de **K** deverão ser arredondados para o maior valor inteiro.*

Referências

AMAZONAS (Estado). **Curso de qualificação profissional**: estatística básica. Amazonas: Manaus, Governo do Estado do Amazonas, 2013. Disponível em: <http://www.seplan.am.gov.br/arquivos/download/arqeditor/apostila_estatistica.pdf>. Acesso em: 3 abr. 2013.

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. **Estatística para cursos de engenharia e informática**. 2. ed. São Paulo: Atlas, 2004.

BRUNI, A. L. **Estatística aplicada à gestão empresarial**. 2. ed. São Paulo: Atlas,. 2008.

GARCIA, R. **Estatística**. São Paulo: Pearson Education do Brasil, 2009.

GUIA do Excel. **Como criar um gráfico de box-plot Excel**. 6 nov. 2012. Disponível em: <http://guiadoexcel.com.br/como-criar-um-grafico-de=box-plot-excel>. Acesso em: 12 maio 2014.

IBGE. Centro de Documentação e Disseminação de Informações. **Normas de apresentação tabular**. 3. ed. Rio de Janeiro: IBGE, 1993.

LARSON, R.; FARBER, B. **Estatística aplicada**. 2. ed. São Paulo: Pearson, 2008.

MCCLAVE. **Estatística para administração e economia**. 10. ed. São Paulo: Pearson Prentice Hall, 2009.

MEDRI, W. **Análise exploratória de dados** — Curso de Especialização “Lato Sensu” em Estatística. Londrina/PR, março de 2011.

MICROSOFT. **Gráfico de linhas (construtor de relatórios e SSRS)**. Disponível: <http://msdn.microsoft.com/pt-br/library/dd207036.aspx>. Acesso em: 12 maio 2014.

SOUZA, A., M. S.; GESSER, K.; DALPIAZ, M. V. A. D. **Caderno de estatística**. São Paulo: Uniasselvi, 2011.

Unidade 3

Cálculo de probabilidades

Ana Luisa Fantini Schmitt

Objetivos de aprendizagem: Esta unidade tem como objetivos: descrever os conceitos fundamentais da teoria das probabilidades; assinalar e classificar variáveis; demonstrar as diferentes aplicações e os tipos de distribuições; utilizar os principais testes estatísticos para análise de amostras e aplicar a inferência em problemas práticos.

Seção 1: **Conceitos fundamentais da probabilidade**

Nesta seção, você aprenderá os conceitos fundamentais da probabilidade, o espaço amostral e o evento. Além disso, o estudo desta seção permitirá a você fazer cálculos de probabilidades em diversas situações aplicadas.

Seção 2: **Teorema de Bayes**

Nesta seção, você aprenderá sobre Thomas Bayes e sua influência para o cálculo de probabilidades por meio do Teorema de Bayes, que facilita o cálculo de probabilidades condicionadas.

Introdução ao estudo

Nesta unidade, você conhecerá o espaço amostral e o evento — conceitos fundamentais da probabilidade — que darão suporte para você fazer cálculos probabilísticos. Esses cálculos são relativamente simples, pois dependem muito da sua interpretação, uma vez que, matematicamente falando, são a razão entre esses dois conceitos (espaço amostral e evento).

A partir deles, você estudará a probabilidade da união de eventos, a probabilidade condicionada e o teorema da multiplicação. Todos esses conceitos estarão sempre permeados por exemplos e resoluções detalhadas, fique tranquilo!

Sabendo que probabilidade está presente em diversas situações, também podemos concluir, desde já, que há várias maneiras de calcular e, consequentemente, prever como os eventos ocorrerão. É neste ponto que você vai conhecer Thamas Bayes e o seu famoso teorema!

O que Bayes fez foi lançar uma certa subjetividade ao prever a ocorrência de um evento. Basta pensar que as informações que você tem sobre as condições de ocorrer o evento vão acabar influenciando a previsão. Por exemplo, grosso modo, você sabe que, em um determinado sorteio, “nunca” saem números entre 25 e 30. Você sabe disso com base em sorteios anteriores e, assim, sabe que as chances desses números saírem em um próximo sorteio serão menores do que as chances dos demais números.

Outro exemplo comum do que propôs Bayes é o possível resultado de um jogo de futebol. Se você quer prever a chance de um time A vencer um time B, deve levar em conta as informações que se tem sobre os resultados das partidas anteriores, como, por exemplo, quantas vezes A venceu B e vice-versa, além de levar em consideração as opiniões de especialistas sobre o jogo, os times, o campeonato e os jogadores.

Com essas informações iniciais, esperamos que você goste do conteúdo que estudará a seguir e que ele lhe sirva para refletir sobre situações, analisar e propor possíveis melhorias em condições cotidianas e profissionais. Bons estudos!

Seção 1 Conceitos fundamentais da probabilidade

É comum fazermos previsões sobre determinados acontecimentos, sobretudo quando já sabemos algo anterior a tal acontecimento. Um jogo de futebol disputado entre os times A e B é um exemplo. Você pode se basear no número de vitórias de A e prever que, por ter ganhado mais vezes e ter obtido melhores resultados do que B, A será vencedor.

Porém, existem fenômenos cujo resultado você não poderá prever mesmo que eles se repitam inúmeras vezes e nas mesmas condições. Um bom exemplo é o lançamento de um dado. Se você jogar um dado honesto (não viciado), jamais poderá saber qual será o próximo resultado antes de lançá-lo.

São resultados como esse, imprevisíveis, que são chamados de **aleatórios**. Para Fonseca e Martins (1996), os fenômenos aleatórios levam a diferentes resultados. Mesmo que se faça o experimento em condições normais e iguais, não há como prever o resultado.

Não há como prever, mas há como palpitar! E justamente para que esse palpite possa ser levado em consideração e seja coerente, os matemáticos criaram a Teoria das Probabilidades.

A probabilidade quantifica a chance de alguma resposta para determinado fenômeno. Essa quantificação será dada em forma de um número entre 0 e 1, em que o 0 representa a resposta impossível no fenômeno realizado, e o 1, a certeza absoluta de que sairá a resposta na próxima jogada.

Um simples exemplo se dá no lançamento de uma moeda. A probabilidade de sair cara é de 0,5, pois metade das opções da moeda (cara ou coroa) representa a resposta esperada, e nenhum lado tem vantagem sobre o outro.

Ao estudar a probabilidade, há dois conceitos importantes a serem entendidos, o espaço amostral e o evento, pois subsidiarão os cálculos iniciais. De certa maneira, a probabilidade é baseada no cálculo da razão entre esses dois conceitos.

1.1 Espaço amostral

O espaço amostral nada mais é do que o conjunto de todas as soluções possíveis dentro de um experimento qualquer, chamado S. O número de soluções possíveis dentro do espaço amostral é chamado n(S). Compreenda essa definição a partir de alguns exemplos:

Exemplo 1:

Ao lançar um dado de seis faces, as soluções possíveis são 1, 2, 3, 4, 5 e 6; sendo assim, determine S e n(S).

S = {1,2,3,4,5,6}, todas as faces possíveis de um dado.

n(S) = 6, pois S possui 6 elementos.

Exemplo 2:

Levando em consideração as vogais do alfabeto, determine S e n(S).

S = {a,e,i,o,u}, ou seja, um conjunto com todas as vogais.

n(S) = 5 porque S possui 5 elementos.

Exemplo 3:

Considerando os naipes de um baralho, determine S e n(S).

S = {espadas, copas, ouros, paus}.

n(S) = 4.

Exemplo 4:

Se você lançar um dado de seis faces duas vezes, qual será o S e o n(S)?

Lembre-se do Princípio Fundamental da Contagem, pois você tem duas etapas com seis maneiras cada uma.

Assim, para cada número que sair no primeiro lançamento, há seis opções de números para o segundo lançamento!

S = {(1,1), (1,2),...,(2,3),...,(3,4),...,(4,5),...(5,6),...,(6,6)}

n(S) = 36.

Exemplo 5:

E se você quiser saber o S e o n(S) que resultam do jogo de seis dezenas feito na mega-sena?

Lembre-se de que você quer saber quantas combinações simples de 6 elementos pode formar com os 60 números. Sendo assim, para determinar o n(S) basta aplicar a fórmula da Combinação:

$C_{60,6}=\frac{60!}{6!54!}=50036860$, ou seja, o n(S) é 50.063.860.

O S seria o conjunto com todas as 50.063.860 possibilidades de jogos!

Para saber mais

Nos exemplos apresentados até aqui, fica fácil determinar o espaço amostral. Porém, há casos em que o espaço amostral tem uma quantidade de elementos muito grande. Em casos como esses, será necessário que você utilize as técnicas de contagem estudadas anteriormente, afinal, para o cálculo de probabilidade, o que interessa é a quantidade de elementos!

Para os cálculos de probabilidades, o n(S) é utilizado; portanto, não se preocupe com o tamanho do espaço amostral, pois em casos como os do Exemplo 5, você não precisaria listar as 50.063.860 combinações para determinar o S!

1.2 Evento

O evento é um subconjunto do espaço amostral, ou seja, uma parte de S. O evento pode ser considerado, também, o conjunto das respostas esperadas, podendo ser representado por uma letra maiúscula (A, B,..., Z).

Nesse caso, o n(A, B,..., Z) identifica o número de respostas esperadas. Compreenda essa definição também com base em alguns exemplos:

Exemplo 1:

Suponha que você lançou um dado e quer saber a probabilidade de sair um número menor ou igual a três.

As soluções são sair o 1, o 2 ou o 3 no lançamento do dado.

Então, o evento pode ser chamado de A = {1,2,3}.

Assim, n(A) = 3, pois há 3 elementos no conjunto A.

Exemplo 2:

E, se você sortear, ao acaso, dentre as 21 consoantes do alfabeto, alguma que está na palavra PROBABILIDADE.

Veja que na palavra PROBABILIDADE existem as consoantes p, r, b, l e d. Então, o evento B será sortear, dentre as consoantes existentes, as consoantes (p, r, b, l, d). Logo, B = { p, r, b, l, d } e n(B) = 5.

Exemplo 3:

Você deve obter um número ímpar, dentre os números de 2 dígitos distintos, que se possa formar com os dez algarismos de 0 a 9.

Para um número ser ímpar, o dígito da unidade deve terminar em 1, 3, 5, 7 ou 9.

Para calcular, pode-se primeiro escolher o dígito da unidade (5 modos) e depois escolher o dígito da dezena (5 modos).

Assim, se o evento fosse identificado como D, n(D) seria 5 . 5 = 25, ou seja, os números ímpares formados. Da mesma forma, D seria o conjunto dos 25 números ímpares formados.

1.3 Probabilidade geral

A probabilidade é uma função que associa um número real, entre 0 e 1, à chance de ocorrência de um evento qualquer A, dentro de um espaço amostral S. A probabilidade de ocorrer o evento A é identificada como P(A) e calculada por

$$P(A)=\frac{n(A)}{n(S)}$$

Onde:

A = conjunto evento.

S = conjunto espaço amostral.

n(A) = número de elementos do conjunto A.

n(S) = número de elementos do conjunto S.

P(A) = probabilidade de ocorrer A.

Existem algumas propriedades relacionadas à probabilidade de A listadas a seguir:

I. $0 \leq P(A) \leq 1$

Se A é subconjunto de S, pode-se afirmar que $n(A) \leq n(S)$, o que implica que P(A) é um número entre 0 e 1.

II. P(S) = 1

Se $P(S)=\frac{n(S)}{n(S)}$, então P(S) = 1.

III Se A = 0, então P(A) = 0.

Se A = 0, tem-se n(A) = 0. Com isso, $P(A)=\frac{n(A)}{n(S)}=\frac{0}{n(S)}=0.$

IV. $P(A)+P(\overline{A})=1$, onde $\overline{A}$ é o conjunto “não A”, chamado de complementar de A.

Veja que, se S é o universo, então os subconjuntos A obedecerão à propriedade: $n(A)+n(\overline{A})=n(S)$.

Então: $P(A)+P(\overline{A})=\dfrac{n(A)}{n(S)}+\dfrac{n(\overline{A})}{n(S)}=\dfrac{n(A)+n(\overline{A})}{n(S)}=\dfrac{n(S)}{n(S)}=1$

Compreenda detalhadamente por meio dos exemplos a seguir:

<u>Exemplo 1</u>:

Determine a probabilidade de sortear, ao acaso, uma carta de copas de um baralho comum.

Solução:

A = sair carta de copas.

A = {copas}.

S = todos os possíveis naipes de um baralho.

S= {copas, espadas, ouros, paus}.

n(A) = 1.

n(S) = 4.

$$P(A)=\frac{n(A)}{n(S)}=\frac{1}{4}=0{,}25\ .$$

Para saber mais

Um baralho comum tem 52 cartas, divididas em 4 naipes (espadas, copas, ouros e paus), e cada naipe possui 13 cartas (2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A).

Para conhecer mais, acesse: <http://www.brasilescola.com/curiosidades/baralho.htm>.

<u>Exemplo 2</u>:

No lançamento de um dado de seis faces, qual a probabilidade de a face que cair voltada para cima ser um número menor ou igual a três?

Solução:

B = sair, no dado, as faces 1, 2 ou 3= {1,2,3}.

S = as faces de um dado = {1,2,3,4,5,6}.

n(B) = 3.

n(S) = 6.

$$P(B)=\frac{n(B)}{n(S)}=\frac{3}{6}=\frac{1}{2}=0{,}5\ .$$

Exemplo 3:

Qual é a probabilidade de que, ao sortear uma letra do alfabeto, esta seja uma vogal “a” ou “i”?

Solução:

C = sortearmos uma vogal = {a, i }.

S = o alfabeto completo = {a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u,w,x,y,z}.

n(C) = 2.

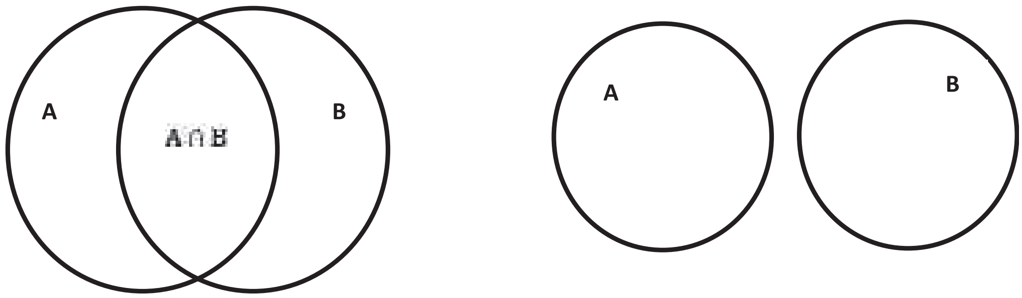
n(S) = 26.

$$P(C) = \frac{n(C)}{n(S)} = \frac{2}{26} = \frac{1}{13} = 0{,}0769 \, .$$

1.3.1 Probabilidade da união de eventos

A união de eventos pode ser caracterizada de duas maneiras, pois existem eventos mutuamente exclusivos e eventos não mutuamente exclusivos. Observe o esquema a seguir que representa essa diferença:

Figura 3.1 Esquema de eventos exclusivos e não mutuamente exclusivos



Fonte: Da autora (2014).

Para identificar se o evento é mutuamente exclusivo ou não mutuamente exclusivo, basta identificar a intersecção entre os eventos. A diferença é que a intersecção pode existir ou não.

Resumindo, eventos mutuamente exclusivos têm intersecção existente; já os não mutuamente exclusivos não têm intersecção existente; nesse caso, ela é representada pelo conjunto vazio.

Segundo Machado (1986), quando A e B são eventos mutuamente exclusivos, A∩B = conjunto vazio — a ocorrência de um deles não implica a não ocorrência do outro.

Exemplo 1: Evento mutuamente exclusivo.

Considerando um baralho completo (52 cartas), qual é a probabilidade de sair uma dama ou um rei?

Eventos:

A = sair dama.

n(A) = 4.

B = sair rei.

n(B) = 4.

A∩B = conjunto vazio, pois não há carta que seja, ao mesmo tempo, dama e rei.

Exemplo 2: Evento não mutuamente exclusivo.

Considerando um baralho completo (52 cartas), qual é a probabilidade de sair uma carta de paus ou um dama?

Eventos:

A = sair qualquer carta de paus.

n(A) = 13.

B = sair qualquer dama.

n(B) = 4.

A∩B = damas de paus.

n(A∩B) = 1, pois existe uma dama de paus no baralho comum.

Agora, pense no que acontece na união entre os eventos A e B nos dois exemplos anteriores. Primeiramente, analisaremos o caso do exemplo 1, no qual a intersecção entre A e B é vazia.

Unindo os eventos A e B (damas ou reis), qual o total de cartas diferentes? A resposta é 8 cartas, 4 damas e 4 reis.

Então, você pode pensar que o número de elementos da união desses dois conjuntos é a soma do número de elementos de cada conjunto: n (A U B) = n (A) + n (B).

Assim, pela demonstração, a probabilidade de sair uma dama ou um rei, ao retirar-se uma carta de um baralho normal, será:

$$P(\mathrm{A} \cup \mathrm{B}) = \frac{n(\mathrm{A} \cup \mathrm{B})}{n(S)} = \frac{n(\mathrm{A})}{n(S)} + \frac{n(\mathrm{B})}{n(S)} = P(\mathrm{A}) + P(\mathrm{B})$$

Ou seja, P(A U B) = P(A) + P(B).

Analisando o exemplo 2, perceba que o conceito acima tem que ser ajustado porque, unindo os eventos A e B (paus e dama), tem-se um total de 16 cartas, mas, se for calculado n(A) + n(B), o resultado será 17.

Parece que está errado? Mas não está! Está correto!

Veja bem, como a intersecção no caso acima é vazia, omitiu-se sua interferência no cálculo. Porém, no exemplo 2, como a intersecção existe, ela acaba interferindo.

Em relação aos conjuntos não mutuamente exclusivos, temos:

n (A U B) = n (A) + n (B) — n (A U B).

Pela demonstração a seguir, a probabilidade entre conjuntos não mutuamente exclusivos é:

$$P(\mathrm{A}\cup\mathrm{B}) = \frac{n(\mathrm{A}\cup\mathrm{B})}{n(S)} = \frac{n(\mathrm{A})+n(\mathrm{B})-n(\mathrm{A}\cap\mathrm{B})}{n(S)} = \frac{n(\mathrm{A})}{n(S)}+\frac{n(\mathrm{B})}{n(S)}-\frac{n(\mathrm{A}\cap\mathrm{B})}{n(S)} =$$

$$P(\mathrm{A}\cup\mathrm{B}) = P(\mathrm{A})+P(\mathrm{B})-P(\mathrm{A}\cap\mathrm{B})$$

Resumindo, a probabilidade da união entre dois eventos A e B será:

P(A U B) = P(A) + P(B) quando A e B são mutuamente exclusivos.

P(A U B) = P(A) + P(B) — P(A ∩ B) quando A e B são não mutuamente exclusivos.

<u>Exemplo 3</u>:

Numa urna, há 10 bolas verdes, 8 azuis e 4 brancas. Qual é a probabilidade de uma bola azul ou branca ser retirada ao acaso?

Solução:

A = sair bola azul.

n(A) = 10, pois há 10 bolas dessa cor na urna.

B = sair bola branca.

n(B) = 4, pois há 4 bolas dessa cor na urna.

S = todas as bolas da urna.

n(S) = 22, pois há 10 verdes, 8 azuis e 4 brancas.

A∩B = bola branca e azul ao mesmo tempo.

n(A∩B) = 0, pois não há bola assim na urna.

Como a intersecção é vazia, os eventos são mutuamente exclusivos. Então, a probabilidade de sair bola azul ou branca é dada pela união dos eventos A e B.

$$P(A \cup B) = P(A) + P(B) = \frac{10}{22} + \frac{4}{22} = \frac{14}{22} = \frac{7}{11} = 0,6364$$

Exemplo 4:

Numa empresa, há 150 funcionários. Foram ofertadas possibilidades de esportes para que eles praticassem uma vez por semana nas quadras da empresa.

Dentre as várias opções de esportes que cada funcionário pode escolher para praticar, 80 escolheram futebol e 25 escolheram basquete. Sabe-se que 10 desses funcionários praticam os dois esportes.

Qual é a probabilidade de sortear, ao acaso, um funcionário dessa empresa que pratica basquete ou futebol?

Solução:

A = pratica futebol. n(A) = 80.

B = pratica basquete. n(B) = 25.

$A \cap B$ = pratica basquete e futebol. $n(A \cap B) = 10$.

S = todos os funcionários da empresa.

n(S) = 150.

Como a intersecção não é vazia, temos um caso de eventos não mutuamente exclusivos. Com isso, a probabilidade da união dos eventos A com B será:

$$P(A) + P(B) - P(A \cap B) = \frac{80}{150} + \frac{25}{150} - \frac{10}{150} = \frac{95}{150} = 0,6334$$

Para saber mais

Se na pergunta do exercício de probabilidade aparecer a conjunção OU entre os eventos, como nos exemplos anteriores, você deverá calcular a união entre os eventos envolvidos.

1.3.2 Probabilidade condicionada

Mas, e se você quiser calcular a probabilidade de ocorrerem eventos seguidos? Por exemplo, e se você jogar mais de uma vez um dado? Como vai calcular a probabilidade?

Nesses tipos de problemas, em que os eventos são repetidos algumas vezes, podem existir casos com reposição ou casos sem reposição. Acompanhe essa diferença e o cálculo para esses casos nos exemplos a seguir.

Exemplo 1:

Considere um lote de 200 peças, no qual 190 são perfeitas e 10 têm defeitos. Qual é a probabilidade de retirar-se, ao acaso, duas peças com defeito desse lote?

A = retirar a primeira peça com defeito.

B = retirar a segunda peça com defeito.

Com reposição:

Você tira a primeira peça e devolve-a ao lote novamente antes de tirar a segunda peça. Observe que $P(\mathrm{A}) = \frac{n(\mathrm{A})}{n(S)} = \frac{10}{200} = 0{,}05$ e, como a peça tirada é reposta no lote, $P(\mathrm{B}) = \frac{n(\mathrm{B})}{n(S)} = \frac{10}{200} = 0{,}05$.

Sem reposição:

A primeira peça retirada do lote não participa do sorteio para a retirada da segunda peça. Sendo assim, P(A) não muda, ou seja, P(A) = 0,05. Mas, como a peça retirada não é reposta, restam 9 peças defeituosas de um total de 199 peças, então: $P(\mathrm{B}) = \frac{n(\mathrm{B})}{n(S)} = \frac{9}{199} \cong 0{,}045$

Exemplo 2:

De um baralho de 52 cartas, foram retiradas, aleatoriamente e sem reposição, duas cartas. Sabendo-se que a primeira carta retirada é de espadas, qual é a probabilidade de a segunda carta ser uma dama vermelha?

Nesse caso, a segunda probabilidade (B) é influenciada pela ocorrência da primeira (A), o que caracteriza uma probabilidade condicionada, chamada P(B/A) (lê-se: probabilidade de B se ocorreu o evento A).

Para calcular P(B/A), basta saber que você está calculando P(B) em relação ao espaço amostral A, e não ao espaço amostral S.

Suponha que ocorreu o evento A, ou seja, já foi feita a primeira etapa do processo (foi retirada uma carta de espadas). Como você vai calcular a probabilidade de ocorrer B, sendo que já ocorreu A, ou seja, P(B/A)?

Basta imaginar que o espaço amostral é o conjunto A e que a única maneira de ocorrer elementos de B no conjunto A é na intersecção dos dois. Então:

$$P(B/A) = \frac{n(A \cap B)}{n(A)} = \frac{\frac{n(A \cap B)}{n(S)}}{\frac{n(A)}{n(S)}} = \frac{n(A \cap B)}{n(A)}$$

Ou seja: $P(B/A) = \frac{n(A \cap B)}{n(A)}$

Analisando a situação do exemplo, tem-se que $P(B/A) = \frac{1}{13}$.

Para chegar ao resultado usando a fórmula $P(B/A) = \frac{n(A \cap B)}{n(A)}$, tem-se:

B: sair dama vermelha (na segunda retirada).

A: sair carta de espadas (na primeira carta).

n(S) = 52.

n(A) = 13.

P(A) = 13/52.

$n(A \cap B) = 1$.

$P(A \cap B) = 1/52$.

$$P(B/A) = \frac{\frac{1}{52}}{\frac{13}{52}} = \frac{1}{13}.$$

Exemplo 3:

Em um grupo de 40 estudantes do Ensino Médio, tem-se as seguintes características:

Quadro 3.1 Idades e gênero dos estudantes

Idade	Sexo		Total
	M	F	
Menos de 15 anos	15	12	27
Mais de 15 anos	5	8	13
TOTAL	20	20	40

Fonte: Da autora (2014).

Qual é a probabilidade de se sortear, ao acaso, dentre os estudantes, um menino, sabendo que ele tem mais de 15 anos?

Solução:

Olhando a tabela, sabemos que a probabilidade é $P(\mathrm{B}) = \frac{5}{20}$, onde B é o evento de ser menino e, nesse caso, o espaço amostral desconsidera as meninas, pois afirma-se que o sorteado é menino!

Também deve ser feito o cálculo de P(B/A), onde A é o evento de ser menino e B é o evento de ter mais de 15 anos:

$$P(\mathrm{B/A}) = \frac{P(\mathrm{A} \cap \mathrm{B})}{P(\mathrm{A})} = \frac{\frac{5}{40}}{\frac{20}{40}} = \frac{5}{20}$$

Veja que, das duas maneiras, o resultado é o mesmo.

1.3.3 Teorema da multiplicação

O Teorema da Multiplicação permite calcular a probabilidade de ocorrência simultânea de dois eventos a partir das probabilidades condicionais.

A partir da fórmula P(B/A) do cálculo da probabilidade condicionada, tem-se:

P(A ∩ B) = P(A).P(B/A) ou **P(A ∩ B) = P(B).P(A/B)**

Levando em consideração o exemplo 1 (das peças) do item anterior, considere um lote de 200 peças no qual 190 são perfeitas e 10 têm defeitos.

Qual é a probabilidade de serem retiradas, ao acaso, duas peças com defeito desse lote?

A probabilidade de retirada, sem reposição, de duas peças com defeito é dada por:

A = retirar a primeira peça com defeito.

B = retirar a segunda peça com defeito.

$$P(\mathrm{A}) = \frac{10}{200} = \frac{1}{20} \text{ e } P(\mathrm{B/A}) = \frac{9}{199}$$

$$P(\mathrm{A} \cap \mathrm{B}) = P(\mathrm{A}).P(\mathrm{B/A}) = \frac{1}{20}.\frac{9}{199} = \frac{9}{3980} \cong 0{,}0023$$

Calcular $P(\text{B/ A}) = \dfrac{9}{199}$ é fácil, pois como a primeira peça retirada tinha defeito, ficaram ainda 9 peças com defeito e 190 peças perfeitas.

No caso específico em que os eventos são independentes (com reposição), no qual o evento anterior não afeta a probabilidade do evento posterior (caso “a” do item 5), temos que P(B/A) = P(B); então, P(A ∩ B) = P(A).P(B).

Para Lipschutz (1993), um evento B é dito independente de um evento A se a probabilidade de B ocorrer não é influenciada pelo fato de A ter ocorrido ou não. Assim, a probabilidade de retirar, com reposição, duas peças com defeito é:

$$P(\text{A}) = \frac{10}{200} = \frac{1}{20} \text{ e } P(\text{B/ A}) = P(\text{B}) = \frac{10}{200} = \frac{1}{20}$$

Assim, $P(\text{A}\cap\text{B}) = P(\text{A}).P(\text{B}) = \dfrac{1}{20}.\dfrac{1}{20} = 0{,}0025$

Nos exemplos a seguir, você poderá identificar outros casos com e sem reposição.

Para saber mais

Fique atento ao uso do ***e*** do ***ou*** para detectar se é ou não um caso de intersecção.

Exemplo 1:

Qual a probabilidade de, em duas retiradas com reposição, saírem duas cartas de copas de um baralho normal?

Solução:

Calcule a probabilidade de sair copas na primeira carta e copas na segunda carta. Lembre-se de que o *e* representa um caso de intersecção!

A = sair a primeira carta de copas.

n(A) = 13, pois há 13 cartas desse naipe no baralho.

B = sair a segunda carta de copas.

n(B) = 13, pois há 13 cartas desse naipe no baralho.

Como o problema é com reposição, o primeiro evento A não influencia o evento secundário B, ou seja, eles são independentes. Veja:

$$P(\text{A}\cap\text{B}) = P(\text{A}).P(\text{B}) = \frac{13}{52}.\frac{13}{52} = \frac{1}{16} = 0{,}0625$$

Exemplo 2:

Qual é a probabilidade de, em duas retiradas sem reposição, saírem duas cartas de copas de um baralho normal?

Solução:

Mesmo que agora o segundo evento B seja influenciado pelo primeiro evento A, pois não há a reposição da primeira carta de copas, o objetivo é calcular a probabilidade de sair copas na primeira carta e sair copas na segunda carta.

Veja que, novamente, tem-se um caso de intersecção de eventos!

A = sair a primeira carta de copas.

n(A) = 13, pois há 13 cartas desse naipe no baralho.

B = sair a segunda carta de copas.

n(B) = 12 e

n(S) = 51, pois o evento é sem reposição, e o evento A influencia no evento B e no espaço amostral S.

$$P(\mathrm{A}\cap\mathrm{B}) = P(\mathrm{A}).P(\mathrm{B}) = \frac{13}{52}.\frac{12}{51} = \frac{156}{2652} = 0{,}0588$$

Exemplo 3:

Uma urna contém 2 bolas pretas e 3 bolas brancas. Qual é a probabilidade de retirar-se duas bolas, ao acaso e sem reposição, e sair uma bola branca e uma preta?

Solução:

Nesse caso, fique atento à ordem em que as bolas estão saindo, pois pode--se obter a bola branca na primeira retirada e a bola preta na segunda retirada ou ao contrário. São dois casos. Veja:

1º caso:

A = sair bola branca na primeira retirada.

B = sair bola preta na segunda retirada.

$$P(\mathrm{A}\cap\mathrm{B}) = P(\mathrm{A}).P(\mathrm{B/\,A}) = \frac{3}{5}.\frac{2}{4} = \frac{6}{20} = 0{,}3$$

2º caso:

B = sair bola preta na primeira retirada.

A = sair bola branca na segunda retirada.

$$P(\mathrm{B}\cap\mathrm{A}) = P(\mathrm{B}).P(\mathrm{A/\,B}) = \frac{2}{4}.\frac{3}{5} = \frac{6}{20} = 0{,}3$$

Juntando os dois casos, devido à conjunção alternativa *ou*, temos:

$$P[(A \cap B) \cup (B \cap A)] = 0{,}3 + 0{,}3 = 0{,}6$$

Para saber mais

É possível chegar à resposta 0,6 multiplicando-se o primeiro caso por 2. Esse número representa a quantidade de combinações possíveis de A e B.

Questões para reflexão

Em algumas áreas em que for atuar como administrador, você, por vezes, deverá ter conhecimento de processos de produção. Para isso, é relevante o conhecimento em probabilidade, uma vez que, para tomar decisões acerca da troca de uma máquina que tem produzido peças defeituosas, por exemplo, você terá de calcular essas chances reais de ela realmente produzir peças com defeito.

Se tiver interesse, acesse o link: <http://www.portal-gestao.com/gestao/item/2555-t%C3%A9cnicas-de-tomada-de-decis%C3%A3o-para-um-ambiente-de-incerteza.html> e veja a relação entre as tomadas de decisão e as incertezas dentro das empresas, e como essa relação pode ter um pouco de probabilidade implícita.

Nesse caso, na sua opinião, ter esse conhecimento probabilístico auxilia nas tomadas de decisão dentro de uma organização? Reflita sobre isso pensando em casos em que os cálculos poderiam ser empregados.

Atividades de aprendizagem

1. Se você tiver em mãos um baralho comum de 52 cartas, determine a probabilidade de, ao retirar aleatoriamente uma carta do baralho, você pegar:

a) Uma carta vermelha.

b) Um rei.

c) Um 6 de espadas.

d) Um 10 vermelho ou um 4 preto.

e) Uma figura (J, Q ou K).

2. Suponha que você lançou um dado equilibrado apenas uma vez. Determine, então, a probabilidade de a face de cima conter:

a) Um 3.

b) Um número par.

c) Um número ímpar.

d) Um número menor que 6.

3. Agora, suponha que você tenha lançado um dado equilibrado duas vezes e tenha observado a face que caiu voltada para cima. Determine:

a) Os elementos do espaço amostral.

b) Os elementos do evento B: sair soma 7 em suas faces.

c) Os elementos do evento C: sair o mesmo número em ambas as faces.

d) Os elementos do evento D: sair soma menor que 10 em suas faces.

e) Os elementos do evento E: sair soma maior que 12 em suas faces.

Seção 2 Teorema de Bayes

Como você estudou na seção anterior, a probabilidade está presente em diversas situações do seu cotidiano, e há diferenciadas maneiras de calcular e prever como os eventos ocorrerão.

Uma das opções de cálculo mais usadas pelos estatísticos nas previsões, porém, é a que foi desenvolvida no século XVIII por Thomas Bayes (1702-1761). A ideia de Thomas Bayes para o cálculo de probabilidades ficou conhecida como Teorema de Bayes.

Com esse teorema, Bayes inseriu uma certa subjetividade na previsão de um evento, ou seja, para ele, a opinião de quem manuseia os números está presente de modo significativo nos cálculos. Essa opinião é baseada nas informações que você tem sobre as condições de ocorrer o evento e, para Bayes, elas certamente vão influenciar na previsão.

Por exemplo, em uma disputa de cara ou coroa, você certamente concorda que a chance de ganhar é de 50%. Mas, se a moeda for jogada três vezes, segundo Bayes, a previsão deve ser ajustada a cada jogada. Imagine que nas duas primeiras jogadas tenha dado “cara”; então, a chance para a terceira jogada não será mais de 50%.

Usar o cálculo da probabilidade é justamente fazer palpites sobre determinados eventos, se vão ocorrer ou não. Apesar de estar baseado rigidamente na lógica e na razão, o Teorema de Bayes passou por várias controvérsias à medida que os estudos sobre probabilidade e estatística evoluíam.

Ainda hoje, Bayes é criticado por incluir a subjetividade no ajuste dos cálculos. Sob o ponto de vista dos matemáticos conservadores e focados na objetividade, basear-se no “achismo” e na intuição não é correto.

Atualmente, o Teorema de Bayes pode ser aplicado em quase todas as áreas do conhecimento, nas pesquisas científicas ou no cotidiano das pessoas. Autoridades de saúde pública não podem deixar de usar os cálculos probabilísticos na previsão de alcance de uma epidemia.

As economias mundiais não vivem mais sem a previsão de inflação, desemprego ou da alta ou baixa da cotação das moedas. Apesar de sempre terem sido criticados, Thomas Bayes e suas ideias continuam desafiando a intuição e o “achismo” nas apostas e palpites diários.

2.1 O Teorema de Bayes

Com base no que você estudou até aqui, lembre-se de que a probabilidade condicional permite obter a probabilidade de ocorrer um evento A, dado que ocorreu o evento B anterior a A.

Porém, você também pode querer saber o contrário, ou seja, a probabilidade de um evento anterior A ter ocorrido, sabendo que um evento posterior B ocorreu.

Sendo assim, o Teorema de Bayes relaciona a probabilidade de ocorrer um evento posterior à probabilidade de um evento anterior ter ocorrido.

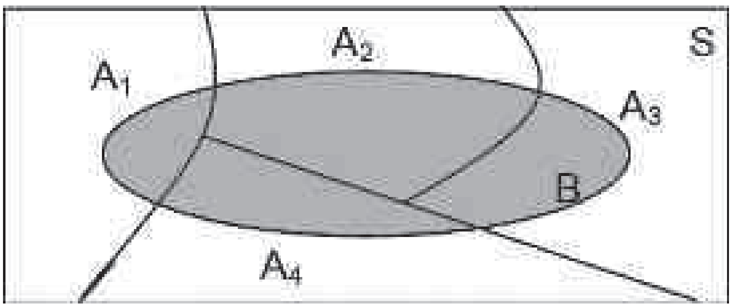
Seja um espaço amostral S subdividido em vários eventos A1, A2, A3,..., Ak que satisfaçam as propriedades:

- $A_i \cap A_j = \varnothing$ para $i \neq j$, em que todos os eventos A_i são mutuamente exclusivos.
- $\cup A_i = S$ em que a união de todos os eventos A_i é igual ao espaço amostral.
- $P(A_i) > 0$ para todo i, em que a probabilidade de qualquer evento A_i é maior que zero.

Além disso, B é um evento qualquer de S.

Observe um caso particular de k = 4 e depois entenda a generalização da ideia.

Figura 3.2 **Generalização do Teorema de Bayes**



Fonte: Da autora (2014).

Observe que a probabilidade de ocorrer o evento B, representado na figura pela área elíptica mais escura, é:

$$\mathbf{P(B) = P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) + P(B \cap A_4)} \text{ (a)}$$

Essa igualdade é válida porque todos os A_i são mutuamente exclusivos dois a dois.

Observe, ainda, que a probabilidade de ocorrer o evento B, dado que ocorreu, por exemplo, o evento A1, é $P(B/A_1) = \frac{P(B \cap A_1)}{P(A_1)}$ ou que a probabilidade de ocorrer B e A1 é $\mathbf{P(B \cap A_1) = P(A_1).\ P(B/A_1).}$

Portanto, a probabilidade de ocorrer B e A_i será:

$\mathbf{P(B \cap A_1) = P(A_1).\ P(B/A_1)}$ (b)

Substituindo (b) em (a), temos:

$$\mathbf{P(B) = P(A_1).\ P(B/A_1) + P(A_2).\ P(B/A_2) + P(A_3).\ P(B/A_3) + P(A_4).\ P(B/A_4)}$$

Para saber a probabilidade de ocorrer, por exemplo, A1, dado que ocorreu B, é:

$$P(A_1/B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{P(A_1).P(B/A_1)}{P(A_1).P(B/A_1) + P(A_2).P(B/A_2) + P(A_3).P(B/A_3) + P(A_4).P(B/A_4)}$$

Generalizando, para k € N, sendo k ≥ 1, tem-se a fórmula do Teorema de Bayes:

$$P(A_i/B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i).P(B/A_i)}{P(A_1).P(B/A_1) + P(A_2).P(B/A_2) + ... + P(A_k).P(B/A_k)}$$

O Teorema de Bayes também revela o Teorema da Probabilidade Total no denominador e representa a probabilidade de ocorrer o evento B.

2.1.1 Exemplos de aplicações do teorema de Bayes

<u>Exemplo 1</u>:

Imagine que você está diante de duas urnas. Uma urna (U1) contém 4 bolas amarelas e 3 pretas. Uma segunda urna (U2) contém 6 bolas amarelas e 2 pretas.

Escolhe-se uma urna ao acaso e dela se retira uma bola, também ao acaso, verificando-se que ela é amarela. Qual é a probabilidade de ela ter vindo da urna U1?

Resolução:

Lembrando que existem duas urnas, considere A o evento: sair bola amarela.

- A probabilidade de escolher uma das urnas é 1/2.
  P(U1) = P(U2) = 1/2.

- A probabilidade de sair bola amarela, dado que se escolheu a U1, é 4/7.
  P(A/U1) = 4/7
- A probabilidade de sair bola amarela, dado que se escolheu a U2, é 6/8.
  P(A/U2) = 6/8 = 3/4

Você quer saber a probabilidade de a bola ter vindo da U1, dado que saiu uma bola amarela, ou seja, P(U1/A) = ?.

Aplique o Teorema de Bayes:

$$P(U_1 / \text{A}) = \frac{P(U_1 \cap \text{A})}{P(\text{A})} = \frac{P(U_1).P(\text{A}/U_1)}{P(U_1).P(\text{A}/U_1) + P(U_2).P(\text{A}/U_2)} =$$

$$P(U_1 / \text{A}) = \frac{\frac{1}{2}.\frac{4}{7}}{\frac{1}{2}.\frac{4}{7} + \frac{1}{2}.\frac{3}{4}} = \frac{\frac{4}{14}}{\frac{4}{14} + \frac{3}{8}} = \frac{\frac{4}{14}}{\frac{74}{112}} = \frac{448}{1036} \cong 0,4324$$

Como a probabilidade de escolher as urnas é igual, a probabilidade de a bola sorteada ter saído da U2 é maior que a probabilidade de ela ter saído da U1, pois a P(A/U2) é maior que a P(A/U1).

<u>Exemplo 2</u>:

Considerando os mesmos números do exemplo 1, qual é a probabilidade de a bola escolhida ser amarela?

Resolução:

A = sair bola amarela.

Você deverá calcular a probabilidade de ocorrer A, mas lembre-se de que A é um conjunto dividido em duas partições do espaço amostral. Há uma parte das bolas brancas em U1 e outra parte em U2. Assim:

$$P(\text{A}) = P(U_1 \cap \text{A}) + P(U_2 \cap \text{A}) = P(U_1).P(\text{A}/U_1) + P(U_2).P(\text{A}/U_2) =$$

$$= \frac{1}{2}.\frac{4}{7} + \frac{1}{2}.\frac{3}{4} = \frac{4}{14} + \frac{3}{8} = \frac{74}{112} \cong 0,6607$$

Veja: o que se fez foi apenas aplicar o Teorema da Probabilidade Total. A resposta é igual ao denominador do Teorema de Bayes do exemplo 1. Confira!

<u>Exemplo 3</u>:

Uma empresa possui 4 máquinas que fazem os mesmos tipos de peças (X, Y, Z) e que se distinguem pela nacionalidade. A probabilidade das peças defeituosas de cada máquina é apresentada no seguinte quadro:

Quadro 3.2 Nacionalidade das máquinas e peças defeituosas

Nacionalidade da máquina	Probabilidade de peças defeituosas		
	X	Y	Z
Americana	0,15	0,21	0,10
Brasileira	0,10	0,20	0,15
Chinesa	0,12	0,18	0,20
Dinamarquesa	0,15	0,11	0,15

Fonte: Da autora (2014).

Escolhendo uma máquina, ao acaso, e retirando aleatoriamente uma peça, verificou-se que é uma peça do tipo Y com defeito. Qual é a probabilidade de essa peça ter sido produzida pela máquina brasileira?

Resolução:

Admitindo que cada máquina tem a mesma chance de ser escolhida, tem-se que: P(A) = P(B) = P(C) = P(D) = 1/4 = 0,25.

A 3ª coluna do quadro anterior indica a probabilidade de a peça Y com defeito ser produzida por uma dada máquina. Então, as probabilidades condicionais da peça Y defeituosa de cada máquina são:

P(Y/A) = 0,21 P(Y/B) = 0,20 P(Y/C) = 0,18 P(Y/D) = 0,11

Assim, a probabilidade de a peça Y ter sido produzida pela máquina brasileira, dado que a peça é defeituosa, é:

$$P(\mathrm{B}/Y) = \frac{P(\mathrm{B}).P(Y/\mathrm{B})}{P(\mathrm{A}).P(Y/\mathrm{A}) + P(\mathrm{B}).P(Y/\mathrm{B}) + P(C).P(Y/C) + P(D).P(Y/D)}$$

$$P(\mathrm{B}/Y) = \frac{0{,}25.0{,}20}{0{,}25.0{,}21 + 0{,}25.0{,}20 + 0{,}25.0{,}18 + 0{,}25.0{,}11} = \frac{0{,}05}{0{,}175} = 0{,}2857$$

Assim, pode-se calcular, da mesma maneira, a probabilidade de a peça Y ter sido produzida em qualquer outra máquina, dado que a peça é defeituosa.

$$P(\mathrm{A}/Y) = \frac{0{,}25.0{,}21}{0{,}25.0{,}21 + 0{,}25.0{,}20 + 0{,}25.0{,}18 + 0{,}25.0{,}11} = \frac{0{,}0525}{0{,}175} = 0{,}30$$

$$P(C/Y) = \frac{0{,}25.0{,}18}{0{,}25.0{,}21 + 0{,}25.0{,}20 + 0{,}25.0{,}18 + 0{,}25.0{,}11} = \frac{0{,}045}{0{,}175} = 0{,}2571$$

$$P(D/Y) = \frac{0{,}25.0{,}18}{0{,}25.0{,}21 + 0{,}25.0{,}20 + 0{,}25.0{,}18 + 0{,}25.0{,}11} = \frac{0{,}0275}{0{,}175} = 0{,}1571$$

Para saber mais

O denominador (0,175), que é o mesmo em todos os casos, representa a probabilidade de escolher, ao acaso, uma peça Y com defeito. Observe que a soma das probabilidades das quatro máquinas é 1.

Analisando as quatro probabilidades calculadas anteriormente, veja que a peça Y com defeito tem mais chance (30%) de ter sido produzida pela máquina americana do que por qualquer uma das outras.

Esse fato também poderia ter sido percebido analisando-se o quadro de probabilidade de peças defeituosas, uma vez que a probabilidade de escolher qualquer máquina é igual, e a probabilidade de a máquina americana produzir uma peça Y defeituosa P(Y/A) é maior do que a das outras máquinas.

Agora, se em vez de a empresa ter comprado a máquina chinesa, tivesse optado por mais uma dinamarquesa, será que a peça Y com defeito teria mais chance de ter sido produzida pela máquina americana? Verifique isso no seguinte exemplo.

Exemplo 4:

Considerando a mesma situação do exemplo 3, mas com duas máquinas dinamarquesas, qual máquina tem mais chance de ter produzido a peça Y com defeito?

Quadro 3.3 Nacionalidade das máquinas e peças defeituosas

Nacionalidade da máquina	Probabilidade de peças defeituosas		
	X	Y	Z
Americana	0,15	0,21	0,10
Brasileira	0,10	0,20	0,15
Chinesa	0,15	0,11	0,15
Dinamarquesa	0,15	0,11	0,15

Fonte: Da autora (2014).

Resolução:

Veja que aumentou a chance de a máquina escolhida ser a dinamarquesa:

$$P(A) = \frac{1}{4}$$

$$P(B) = \frac{1}{4}$$

$$P(D) = \frac{1}{2}$$

As probabilidades condicionais continuam as mesmas:

P(Y/A) = 0,21 P(Y/B)= 0,20 P(Y/D) = 0,11

Calculando a probabilidade de a peça Y ter sido produzida por cada uma das máquinas, dado que a peça é defeituosa:

$$P(A/Y) = \frac{0,25.0,21}{0,25.0,21 + 0,25.0,20 + 0,25.0,18 + 0,5.0,11} = \frac{0,0525}{0,1575} \cong 0,3334$$

$$P(B/Y) = \frac{0,25.0,18}{0,25.0,21 + 0,25.0,20 + 0,25.0,18 + 0,5.0,11} = \frac{0,045}{0,1575} \cong 0,3175$$

$$P(D/Y) = \frac{0,25.0,18}{0,25.0,21 + 0,25.0,20 + 0,25.0,18 + 0,5.0,11} = \frac{0,055}{0,1575} \cong 0,3492$$

Apesar de a máquina americana ter maior probabilidade de produzir uma peça Y defeituosa, a peça Y com defeito tem mais chance (34,92%) de ter sido produzida pela máquina dinamarquesa.

Isso não quer dizer que o fato de ter mais máquinas do tipo dinamarquesa sempre contribuirá para que isso aconteça, pois, se a probabilidade de a máquina dinamarquesa produzir uma peça Y defeituosa P(Y/D) fosse de 0,10, a peça Y com defeito teria mais chance de ter sido produzida pela máquina americana.

96928-978-85-68075-12-8 METODOS_QUATITATIVOS_ALTA_25_06_2014.pdf, page 41 @ Preflight - 14:16:30 - June 30, 2014 - PG-134

Questões para reflexão

Você já parou para pensar que a probabilidade de algo ocorrer ou não faz parte da nossa vida diária?

Por exemplo, qual é a probabilidade de você encontrar um amigo que não vê há anos? Qual é a chance de você investir em algum negócio e ele dar resultado já no primeiro ano? Qual é a probabilidade de você encontrar o amor da sua vida nos corredores da faculdade? Ou mesmo, qual é a chance de você ganhar na loteria? Claro, todos pensamos nessas chances ou probabilidades não é mesmo?

Para ajudá-lo, acesse o link http://economia.ig.com.br/fisico-usa-probabilidades-para-mostrar-como-o-acaso-influencia-o-mundo/n1597200937269.html e conheça o que o físico norte-americano Leonard Mlodinow, de 56 anos, acredita e defende. Aliás, ele mesmo se intitula fruto do acaso, ou da sorte. Afinal, segundo o físico, ele nasceu porque seu pai sobreviveu à 2ª Guerra Mundial e sua vida sempre foi cheia de acontecimentos probabilísticos.

Mas, na sua opinião, isso pode ser utilizado a nosso favor? Realmente, a matemática faz sentido para acontecimentos tão subjetivos? E o acaso? Ele realmente existe? Reflita, pense, e procure equilibrar a importância da probabilidade numérica com o acaso. Quem sabe não seria esse o segredo dos nossos passos futuros... Quem sabe?!

Atividades de aprendizagem

1. Segundo um professor de probabilidade, a chance de um aluno se dar bem em uma prova é de 80% se estudou e de 50% se não estudou. Se um determinado aluno não estuda para 15% das provas que realiza, qual será a probabilidade de esse aluno se dar bem na prova?
2. Considerando os dados da questão anterior, qual a probabilidade de o aluno ter estudado, dado que ele se deu bem na prova?

5 25 50 75 100

Fique ligado!

- Espaço Amostral (chamado S) é o conjunto de todas as soluções possíveis relativas a um experimento qualquer.
- Evento é uma parte do Espaço Amostral.
- A probabilidade de ocorrer um evento A é identificada como:

$$P(A)=\frac{n(A)}{n(S)}=\frac{\text{número de resultados associados ao evento A}}{\text{número de resultados possíveis}}$$

- A probabilidade da união entre dois eventos A e B será:
  $P(A \cap B) = P(A) + P(B)$, se A e B são mutuamente exclusivos.
  $P(A \cap B) = P(A) + P(B) – P(A \cap B)$, se A e B são mutuamente exclusivos.
- A probabilidade de ocorrer o evento B, dado que ocorreu o evento A, é:

$$P(A/B)=\frac{P(A \cap B)}{P(A)}$$

- A probabilidade de ocorrerem os eventos A e B é:
  $P(A \cap B) = P(A) \cdot P(B/A)$, se A e B são eventos dependentes
  $P(A \cap B) = P(A) \cdot P(B)$, se A e B são eventos independentes
- Teorema de Bayes
  A probabilidade de ocorrer um anterior evento Ai, dado que ocorreu um evento posterior B, é:

$$P(A_i/B) = \frac{P(A_i)\cdot P(B/A_i)}{P(A_1)\cdot P(B/A_1) + P(A_2)\cdot P(B/A_2) + ... + P(A_k)\cdot P(B/A_k)}$$

  - Teorema da Probabilidade Total (denominador do Teorema de Bayes)
    A probabilidade de ocorrer o evento B é:
    $P(A_1) \cdot P(B/A_1) + P(A_2) \cdot P(B/A_2) + ... + P(A_k) \cdot P(B/A_k)$

Para concluir o estudo da unidade

Nesta unidade, você pôde conhecer os conceitos básicos da probabilidade, o espaço amostral e o evento. Com isso, fez a relação de que temos uma parte (evento) de um todo (espaço amostral) e, com esses dois conceitos, podemos calcular qualquer probabilidade, seja em união de eventos ou em situações de eventos condicionados.

Outro conteúdo que você estudou foi o Teorema de Bayes. Primeiro, você conheceu o próprio Thomas Bayes e sua trajetória no estudo e aprimoramento das probabilidades. Depois, teve contato com o teorema, que trata da probabilidade de ocorrer um evento dado que ocorreu um evento posterior.

Por exemplo, em uma empresa que fabrica peças, você foi fazer uma inspeção nas máquinas de diversas marcas e, escolhendo uma máquina ao acaso e retirando aleatoriamente uma peça, você percebeu que é uma peça com defeito. Qual é a probabilidade de essa peça ter sido produzida pela máquina da marca X? Com isso, você poderia identificar que a máquina da marca X não é boa, pois produz peças defeituosas em números maiores do que a máquina da marca Y, por exemplo.

Enfim, a probabilidade serve para auxiliar na tomada de decisões. Como futuro administrador, você passará por situações em que o raciocínio matemático e probabilístico será necessário para saber que decisão tomar. Esperamos que o estudo desta unidade tenha lhe permitido perceber isso e que você possa realmente pôr em prática o aprendizado. Desejamos um excelente curso e que você seja um profissional brilhante e de sucesso!

Atividades de aprendizagem

1. No famoso jogo de loteria "mega-sena", são sorteadas 6 dezenas. Você pode montar seu jogo escolhendo algumas entre as 60 dezenas. Supondo que você escolheu 10 dezenas, qual é a probabilidade de você acertar:

a 4 dezenas?

b) 5 dezenas?

c) 6 dezenas?

2. Você jogou dois dados perfeitos e pretende analisar a soma das faces voltadas para cima. Qual é a probabilidade de essa soma ser 6?

3. Em um sorteio de um número natural de 1 a 150, qual é a probabilidade de sair um número múltiplo de 10 ou de 15?

4. No lançamento simultâneo de um dado honesto e uma moeda honesta, qual é a probabilidade de se obter um 5 ou uma cara?

5. Uma urna contém 9 bolas verdes numeradas de 1 a 9 e 6 bolas azuis numeradas de 10 a 15. Se você, ao acaso, retirar uma das bolas, qual será a probabilidade:

a) De sair uma bola verde?

b) De sair uma bola com número ímpar?

c) De sair uma bola azul com número par?

6. Uma máquina produziu 200 peças, das quais 25 estavam com defeito. Ao retirar, aleatoriamente, 4 peças, com reposição, qual é a probabilidade de que:

a) Todas sejam perfeitas?

b) Todas tenham defeito?

c) Pelo menos uma seja perfeita?

d) Pelo menos uma tenha defeito?

7. Numa sala de aula com 28 estudantes, 15 são meninos e 12 são morenos, dos quais 7 são meninas. Escolhendo um estudante ao acaso, qual é a probabilidade de ele ser moreno ou menina?

8. A probabilidade de que o filho de um casal nasça com olhos verdes é de $\frac{1}{4}$. Se o casal tiver três filhos, qual é a probabilidade de todos terem olhos verdes? E de nenhum ter olhos verdes? E de, pelo menos um ter olhos verdes?

9. Um grupo de 83 estudantes apresenta, de acordo com o sexo e a altura, a seguinte situação:

Quadro 3.4 Altura dos estudantes

	Menos de 1,40m	De 1,40m a 1,60m	Mais de 1,60m
Meninos	15	12	8
Meninas	11	24	13

Fonte: Do autor (2014).

Escolhendo uma pessoa desse grupo ao acaso, determine:

a) A probabilidade de ser um menino.

b) A probabilidade de ser uma menina com menos de 1,40m.

c) A probabilidade de ser um menino, se o escolhido tiver de 1,40m a 1,60m.

d) A probabilidade de ser uma menina, se o escolhido tiver no máximo 1,60m.

10. Uma prova é composta por 15 questões, e cada uma possui 4 alternativas, das quais apenas uma é correta. Para alguém que esteja respondendo aleatoriamente uma alternativa em cada questão, qual é a probabilidade de:

a) Acertar as 15 questões?

b) Errar as 15 questões?

c) Acertar $\frac{2}{3}$ das questões?

11. Uma caixa em forma de urna contém 5 bolas azuis e 3 rosas. Uma segunda caixa contém 3 bolas azuis e 2 rosas. Se uma bola branca é sorteada ao acaso, qual é a probabilidade de ela ter vindo da primeira caixa?

12. Três empresas de brinquedos A, B e C, produziram, respectivamente, 40%, 50% e 10% do total de brinquedos de uma escola. A porcentagem de brinquedos defeituosos da fábrica A é de 3%, da fábrica B é de 5% e da fábrica C é de 2%. Uma criança recebeu, ao acaso, um brinquedo defeituoso. Qual é a probabilidade de que essa peça tenha vindo da fábrica B? Qual das empresas tem mais chance de ter fabricado a peça defeituosa?

13. Considerando os dados da questão anterior, determine qual é a probabilidade de uma criança receber um brinquedo defeituoso.

14. Considere a existência de três urnas. A primeira urna contém 9 bolas: 3 pretas, 1 branca e 5 vermelhas; a segunda contém 9 bolas: 4 bolas pretas, 3 brancas e 2 vermelhas; finalmente, a terceira urna contém 8 bolas: 2 pretas, 3 brancas e 3 vermelhas. Escolheu-se uma urna, ao acaso, e dela se extraiu uma bola ao acaso. Verificando-se que a

bola sorteada é branca, qual é a probabilidade de a bola ter saído da terceira urna?

15. O proprietário de uma facção estima que uma roupa feita por um funcionário experiente tem 90% de chance de não apresentar defeito, e que uma roupa feita por um funcionário novato tem 50% de chance de não apresentar defeito. Se uma roupa selecionada ao acaso apresentar defeito, determine a probabilidade de ela ter sido feita por um funcionário novato, sabendo-se que 2/3 das roupas são feitas por funcionários experientes.

Referências

BARBETTA, P. A. **Estatística aplicada às ciências sociais**. 3. ed. Florianópolis: Ed. da UFSC, 1999.

BENNETT, D. J. **Aleatoriedade**. Traduzido por Waldéa Barcellos. São Paulo: Martins Fontes, 2003.

FAVARO, S.; KMETEUK FILHO, O. **Noções de lógica e matemática básica**. Rio de Janeiro: Editora Ciência Moderna, 2005.

FILHO, E. de A. **Iniciação à lógica matemática**. São Paulo: Nobel, 2006.

FONSECA, J. S.; MARTINS, G. de A. **Curso de estatística**. 6. ed. São Paulo: Atlas, 1996.

GUERRA, M. J.; DONAIRE, D. **Estatística indutiva**. 5. ed. São Paulo: Livraria Ciência e Tecnologia, 1991.

HEATH, O. V. S. **A Estatística na pesquisa científica**. São Paulo: E. P. U./ EDUSP, 1989.

LIMA, El. et al. **A matemática do ensino médio**. 5. ed. Rio de Janeiro: Sociedade Brasileira de Matemática, 2004. v. 2.

LIPSCHUTZ, S. **Probabilidade**. Traduzido por Ruth Ribas Itacarabi. 4. ed. São Paulo: Makron Books, 1993.

MACHADO, A. dos S. **Matemática, temas e metas**: sistemas lineares e análise combinatória. São Paulo: Atual, 1986.

MAGALHÃES, M. N.; LIMA, A. C. P. de. **Noções de probabilidade e estatística**. São Paulo: Universidade de São Paulo, 2004.

MORETTIN, L. G. **Estatística Básica**: probabilidade. São Paulo: Makron, 1999. v. 1.

SPIEGEL, M. R.; SCHILLER, J.; SRINIVASAN, A. **Teoria e problemas de probabilidade e estatística**. Porto Alegre: Bookman, 2004.

Unidade 4

Cálculos estatísticos

Débora Cristina Brandt

Objetivos de aprendizagem: Nesta unidade, você será levado a conhecer, diferenciar e calcular a média simples e a média ponderada, bem como realizar a regressão linear a partir do método de mínimos quadrados.

Seção 1: **Média e média ponderada**

Tanto a média como a média ponderada nos fornecem uma medida resumo dos dados obtidos em uma pesquisa estatística. Veremos nesta seção qual a diferença entre os dois e aprenderemos como calculá-los.

Seção 2: **Variância e desvio-padrão**

A variância é uma medida de dispersão de dados de uma pesquisa estatística em torno da média. Nesta seção, iremos aprender a calculá-la, assim como, também, aprenderemos a calcular o desvio-padrão.

Seção 3: **Correlação e regressão linear**

A correlação é uma medida que mede a dependência linear entre duas variáveis em uma mesma pesquisa estatística. Se houver dependência, poderemos encontrar a equação da reta que melhor descreva essa situação através da regressão linear.

Introdução ao estudo

Nas unidades anteriores, você aprendeu a apresentar os dados referentes às variáveis estatísticas por meio de tabelas e gráficos. Embora essa representação ajude a tirar conclusões sobre o que foi pesquisado, existem outras maneiras complementares a elas que permitem extrair informações: são os chamados cálculos estatísticos.

Quando estudamos uma variável em particular, podemos resumir as informações apresentadas em uma tabela por média e da variância dos dados. Assim, conseguimos expressar através de dois números o comportamento da variável em estudo.

No caso de estudarmos tabelas de dupla entrada, onde estamos interessados não só no comportamento da variável, mas na maneira como uma influencia o comportamento da outra, podemos calcular a correlação linear entre elas, isto é, é possível quantificar a dependência linear entre duas variáveis. Se elas estiverem correlacionadas linearmente, podemos representá-las por meio de um gráfico e aproximá-las utilizando para isso a regressão linear. Serão esses conceitos que estudaremos nesta unidade.

Na primeira seção, apresentaremos os conceitos de média e média ponderada, qual a diferença entre os dois e como calculá-los. Na segunda seção, aprenderemos a calcular a variância e o desvio-padrão de uma variável, que quantifica a dispersão dos valores observados em torno da média. Já na terceira seção, estudaremos a correlação linear entre duas variáveis e a maneira de relacioná-las via regressão linear.

Seção 1 Média e Média Ponderada

Suponha que você está indo, pela primeira vez, consultar determinado dentista. Ao chegar lá, observa que, embora seu horário esteja se aproximando, há uma quantidade razoável de pessoas na sala de espera — ou seja, a consulta irá atrasar. Como você é uma pessoa ocupada e sabe que vai ter que voltar lá inúmeras vezes, gostaria de ter uma ideia do quanto as consultas costumam atrasar. Então, você resolve perguntar para a secretária a respeito. Ao perguntar isso, nem passa pela sua cabeça que a secretária vai lhe fornecer uma lista com todos os atrasos para aquele horário no último ano ou mês. Na verdade, ela vai

lhe dar uma única informação que vai fazer com que você tenha sua pergunta razoavelmente respondida. Essa é a ideia das medidas de posição, ou medidas de tendência central para um conjunto de dados qualquer.

Suponhamos que o dentista também esteja atento a essa questão e resolveu pedir para sua secretária anotar a quantidade de minutos que cada paciente tem que esperar pra ser atendido. No dia anterior à sua consulta, por exemplo, foram atendidas 20 pessoas, e os atrasos observados (em minutos) foram os seguintes:

10	15	8	15	22	15	30	21	15	18
33	42	45	22	15	18	22	25	18	22

Vamos partir desses valores para entender o significado de cada medida de posição que apresentaremos a seguir.

1.1 Média aritmética

A média aritmética é uma medida de posição para variáveis quantitativas e é obtida somando-se todos os valores observados e dividindo-se o resultado pelo número de observações. Formalmente, se X for uma variável com observações $x_1, x_2, ..., x_n$, chamamos de média de X à soma dos valores dividida pelo número de observações, ou seja,

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

Vamos calcular o tempo médio de espera no dentista?

A variável em questão é o tempo de espera, em minutos, e o número de observações é n = 20. Assim,

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{10 + 15 + 8 + 15 + 22 + ... + 22 + 25 + 18 + 22}{20} = \frac{431}{20} = 21{,}55$$

Portanto, o tempo médio de espera no dentista naquele dia foi de 21,55 min.

Observe que o valor encontrado para a média não foi observado: de acordo com os dados obtidos pela secretária, ninguém esperou 21,55 min naquele dia. Na verdade, o valor médio não precisa ser igual a um dos dados observados na pesquisa.

1.2 Média ponderada

Para calcularmos o tempo médio de atraso, realizamos uma soma com 20. Uma maneira mais eficiente de calcular a média é através da distribuição de frequências. Nesse caso, poderemos agrupar os dados que aparecem mais de uma vez.

Vamos montar a tabela para o nosso exemplo:

Tabela 4.1 Exemplo de média aritmética

Atraso no dentista no dia 'a' - 2014

Tempo (min)	ni
8	1
10	1
15	5
18	3
21	1
22	4
25	1
30	1
33	1
42	1
45	1
Total	**20**

Fonte: Dados fictícios.

Assim, em vez de somarmos 20 + 20 + 20 + 20, fazemos 4×20. E, dessa forma, em vez de realizarmos uma soma com 20 parcelas, faremos uma soma com 11 parcelas.

$$\overline{x} = \frac{1 \cdot 8 + 1 \cdot 10 + 5 \cdot 15 + 3 \cdot 18 + ... + 1 \cdot 45}{20} = \frac{431}{20} = 21{,}55$$

Damos a essa maneira de calcular média o nome de média aritmética. Formalmente, se X for uma variável com observações x_1, x_2, ..., x_l, cujas frequências observadas são n_1, n_2, ..., n_k, respectivamente, com $k \leq n$, calculamos a média de X como:

$$\overline{x} = \frac{n_1 \cdot x_1 + n_2 \cdot x_2 + ... + n_k \cdot x_k}{n_1 + n_2 + ... + n_k},$$

ou, ainda,

$$\overline{x} = \frac{n_1 \cdot x_1 + n_2 \cdot x_2 + ... + n_k \cdot x_k}{n}.$$

Questões para reflexão

Por que as duas equações propostas para a média ponderada são equivalentes?

Exemplo 2: Considere a seguinte tabela estatística:

Tabela 4.2 **Exemplo de média ponderada**

Número de vestibulares prestados antes da aprovação

Número de vestibulares	Ni
0	28
1	54
2	40
3	10
4	8
Total	140

Fonte: Dados fictícios.

Vamos calcular a quantidade média de vestibulares prestados.

$$\overline{x} = \frac{n_1 \cdot x_1 + n_2 \cdot x_2 + ... + n_k \cdot x_k}{n_1 + n_2 + ... + n_k} = \frac{0 \cdot 28 + 1 \cdot 54 + 2 \cdot 40 + 3 \cdot 10 + 4 \cdot 8}{140} = \frac{196}{140} = 1{,}4$$

Note que, se os dados não tivessem agrupados, o trabalho seria razoável, pois o número de dados observados é de 140. Entretanto, como os dados estão agrupados na tabela de frequência, temos uma soma com 5 parcelas.

Vamos criar uma coluna auxiliar na tabela para facilitar o trabalho. Observe:

Tabela 4.3 **Tabela auxiliar para o cálculo da média ponderada**

xi	**ni**	**$x_i \cdot n_i$**
0	28	0·28 = 0
1	54	1·54 = 54
2	40	2·40 = 80
3	10	3·10 = 30
4	8	4·8 = 32
Total	**140**	**196**

Fonte: Dados fictícios.

Assim, calcular a média resume-se a tomarmos a razão 196/140 = 1,4.

E se a variável for quantitativa contínua?

Se os dados não estiverem agrupados, basta aplicar a definição de média: somamos todos os dados observados e dividimos pela quantidade de dados observados. Mas e se os dados estiverem agrupados em uma distribuição de frequência?

Exemplo: Considere a tabela de distribuição de frequência a seguir.

Tabela 4.4 **Distribuição de intervalos de classe**

Altura dos estudantes da classe 2013

	ni
128 |— 131	3
131 |— 134	7
134 |— 137	5
137 |— 140	4
140|— 144	1
Total	20

Fonte: Dados fictícios.

Sabemos que há 3 estudantes na classe com altura entre 128 cm e 131cm, mas não sabemos quantos medem, exatamente, 128 cm ou 129 cm. Para calcularmos a média, nesse caso, precisamos eleger um representante para cada classe, um número que utilizaremos para realizar os cálculos — no caso, o ponto médio do intervalo (URBANO, 2010, p. 98). Depois, basta proceder como anteriormente.

Tabela 4.5 **Tabela auxiliar**

	xi	ni	$x_i \cdot n_i$
128 |— 131	129,5	3	388,5
131 |— 134	132,5	7	927,5
134 |— 137	135,5	5	677,5
137 |— 140	138,5	4	554
140|— 144	141,5	1	141,5
Total		**20**	**2689**

Logo, a altura média dos estudantes da classe em 2013 foi de 2989/20 = 134,45 cm.

A média ponderada pode ser aplicada também quando queremos dar pesos diferentes a valores.

Exemplo: na composição de uma nota semestral, o professor responsável quer realizar 3 provas, mas quer que a primeira tenha peso 1 e a segunda e a terceira tenham peso 2. Se considerarmos N_1, N_2, N_3 como sendo as notas da primeira, da segunda e da terceira avaliações respectivamente, o cálculo da média ponderada será:

$$\overline{x} = \frac{1 \cdot N_1 + 2 \cdot N_2 + 2 \cdot N_3}{(1+2+2)} = \frac{1 \cdot N_1 + 2 \cdot N_2 + 2 \cdot N_3}{5}$$

Suponhamos que um estudante tenha tirado 3, 7 e 8 na primeira, segunda e terceira prova respectivamente. Sua média será, então,

$$\overline{x} = \frac{1 \cdot 3 + 2 \cdot 7 + 2 \cdot 8}{5} = \frac{1 \cdot 3 + 2 \cdot 7 + 2 \cdot 8}{5} = \frac{33}{5} = 6,6$$

Questões para reflexão

Qual a interpretação que pode ser dada aos pesos das notas em termos de tabela de frequências?

Atividades de aprendizagem

1. Um banco instalou um caixa eletrônico em um posto de combustível e está observando o número de usuários que vêm utilizando o serviço. Diariamente, o número de clientes que utilizou o serviço nos últimos 32 dias foi:

15	17	16	15	17	14	17	16	16
17	15	18	14	17	15	14	15	14
15	16	17	18	18	17	15	16	14
18	18	16	15	14				

Qual o número médio de clientes que utilizaram o serviço no período?

2. Exercício adaptado de (MAGALHÃES, 2010): O valor médio de comercialização da saca de milho de 60 quilos na BM&F é apresentado abaixo, em reais, para os últimos 40 meses.

6,1	6,2	6,7	6,5	6,9	6,3	7,4	7,6	7,7	7,6
7,3	7,7	7,6	7,4	7,2	7,2	7,3	7,6	7,5	7,4
7,5	7,7	8,2	8,3	8,1	8,1	8,1	7,9	7,8	7,4
7,5	7,6	7,5	7,6	7,4	7,3	7,4	7,5	7,5	7,4

Organize os dados em faixas de tamanho 0,4 a partir de 6. Calcule o valor médio de comercialização no período.

Para saber mais

É muito importante que saibamos interpretar os dados estatísticos que nos cercam. Interpretações errôneas podem nos levar a conclusões erradas e, até mesmo, absurdas. Nesse sentido, sugerimos o texto de autoria do prof. dr. Marcelo Menezes Reis, que trata justamente de como interpretar os dados estatísticos que nos cercam. Você pode acessá-lo em: http://www.inf.ufsc.br/~marcelo/contest.html>.

Seção 2 Variância e desvio-padrão

Vocês já devem ter ouvido falar no bairro do Morumbi, em São Paulo. Esse bairro é considerado de classe média alta: o Jóquei Club fica no Morumbi, o Palácio dos Bandeirantes — sede do governo do estado —, o Shopping Cidade Jardim, que é um dos mais caros — senão *o* mais caro — shopping da cidade. Mas nesse bairro também fica a maior favela de São Paulo: a Paraisópolis.

Se fosse feita uma pesquisa sobre a renda dos paulistanos de acordo com o bairro em que eles moram, possivelmente a renda média dos moradores do Morumbi seria muito boa, mas será que seria representativa?

Muitas vezes, as medidas de posição por si só não nos dão a informação completa e escondem discrepâncias que deveriam ser conhecidas. Por essa razão, as medidas de posição precisam ser complementadas pelas medidas de dispersão, que nos dizem como os valores se distribuem em torno das medidas de posição.

Assim como no caso das medidas de posição, existem várias medidas de dispersão. Nesta unidade, estudaremos a variância e o desvio-padrão.

2.1 Variância e desvio-padrão populacional

Segundo Magalhães (2010), se X é uma variável com observações x_1, x_2, ..., x_n, chamamos de variância populacional desse conjunto de observações a seguinte equação:

$$Var(X) = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... + (x_n - \overline{x})^2}{n}$$

Note que, para calcular a variância, é necessário como primeiro passo obter a média.

Exemplo: Suponhamos que os dados a seguir são relativos ao tempo de espera em minutos para o atendimento médico em um consultório em certo dia A.

20 30 15 40 38 35 20 24

Inicialmente, vamos calcular a média de tempo de espera:

$$\overline{x} = \frac{20+30+15+40+38+35+20+24}{8} = 27,75 \min$$

De posse da média, podemos calcular a variância:

$$Var(X) = \frac{(20\text{-}27{,}75)^2 + (30\text{-}27{,}75)^2 + (15\text{-}27{,}75)^2 + (40\text{-}27{,}75)^2}{8} +$$
$$+ \frac{(38\text{-}27{,}75)^2 + (35\text{-}27{,}75)^2 + (20\text{-}27{,}75)^2 + (24\text{-}27{,}75)^2}{8}$$
$$= 76{,}19$$

O fato de termos elevado as diferenças ao quadrado faz com que nosso resultado seja dado em $(\text{min})^2$. Se tomarmos a raiz desse valor, voltaremos a ter um número em minutos. Esse procedimento nos dá o que chamamos de desvio-padrão:

$$dp(X) = \sqrt{Var(X)}$$

No nosso caso, o valor do desvio-padrão será $dp(X) = \sqrt{76{,}19} = 8{,}73$ min.

Assim, o tempo médio de espera foi de 27,75min com um desvio-padrão de 8,73min.

O que isso significa?

Significa que o tempo médio de espera é de 27,75min e que os outros tempos de espera não diferem mais do que 8,73min desse valor.

Vimos que, quando a variável está sendo apresentada em uma tabela de frequências, o cálculo da média é facilitado. A mesma coisa acontece com a variância.

Exemplo: Considere os dados a seguir, relativos a uma pesquisa sobre determinada variável quantitativa.

Tabela 4.6 **Exemplo de variância populacional**

Xi	ni
0	7
1	3
2	4
3	2
4	0
5	5
Total	**21**

Fonte: Dados fictícios.

Primeiro, precisamos calcular a média dessa variável fictícia. Para isso, podemos utilizar uma tabela auxiliar:

Tabela 4.7 **Tabela Auxiliar**

Xi	**Ni**	$x_i \cdot n_i$
0	7	0·7 = 0
1	3	1·3 = 3
2	4	2·4 = 8
3	2	3·2 = 6
4	0	4·0 = 0
5	5	5·5 = 25
Total	**21**	**42**

Fonte: Dados fictícios.

Assim, a média será $\overline{x} = \frac{42}{21} = 2.$

Agora, precisamos calcular a variância através da fórmula $Var(X) = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... + (x_n - \overline{x})^2}{n}$. Novamente, uma tabela auxiliar tornará a realização desse cálculo mais simples:

Tabela 4.8 **Tabela Auxiliar**

xi	**ni**	$x_i \cdot n_i$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
0	7	0 · 7 = 0	-2	4
1	3	1 · 3 = 3	-1	1
2	4	2 · 4 = 8	0	0
3	2	3 · 2 = 6	1	1
4	0	4 · 0 = 0	2	4
5	5	5 · 5 = 25	3	9
Total	**21**	**42**		**19**

Assim, $Var(X) = \frac{19}{21} = 0,905$ e, portanto, o desvio-padrão é dado por $dp(X) = \sqrt{0,905} = 0,951$.

2.2 Variância e desvio-padrão amostral

A fórmula que vimos anteriormente é utilizada para o cálculo da variância e do desvio-padrão populacional, isto é, quando toda a população foi considerada. Como normalmente trabalhamos com amostra, precisamos fazer um ajuste nos cálculos por questões técnicas que fogem ao escopo deste caderno de estudos (para saber mais, consulte (MAGALHÃES, 2010). Assim, a **variância amostral** de um conjunto de observações x_1, x_2, ..., x_n é dada pela seguinte equação:

$$S^2(X) = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... + (x_n - \overline{x})^2}{n-1}$$

Observe que a única diferença entre variância populacional e amostral está no denominador da fórmula.

Exemplo: Se os dados a seguir são relativos ao tempo de espera em minutos para o atendimento médico em um consultório para uma amostra de pacientes (não todos) em certo dia A,

20 30 15 40 38 35 20 24

a variância será dada por:

$$S^2 = \frac{(20\text{-}27{,}75)^2 + (30\text{-}27{,}75)^2 + (15\text{-}27{,}75)^2 + (40\text{-}27{,}75)^2}{8-1} +$$
$$+ \frac{(38\text{-}27{,}75)^2 + (35\text{-}27{,}75)^2 + (20\text{-}27{,}75)^2 + (24\text{-}27{,}75)^2}{7}$$
$$= 87{,}07$$

O cálculo do desvio-padrão não muda:

$$S(X) = \sqrt{S^2(X)}$$

No nosso caso, o valor do desvio-padrão será $s(X) = \sqrt{87{,}07} = 9{,}33$ min.

Exemplo: Vamos calcular a variância amostral da tabela a seguir, utilizando uma tabela auxiliar.

Tabela 4.9 **Cálculo da variância amostral.**

Xi	fi
1	6
2	0
3	6
Total	12

Fonte: Dados fictícios

Tabela 4.10 **Utilizando a tabela auxiliar**

xi	fi	$x_i \cdot n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	6	6	-1	1
2	0	0	0	0
3	6	18	1	1
Total	12	24		2
Média $\bar{x}$		2		
Variância Amostral				0,182
Desvio-padrão amostral				0,426

Encontramos a média $\bar{x} = \frac{24}{12} = 2$, a variância amostral $S_2(X) = \frac{2}{(12-1)} = 0{,}182$ e, finalmente, o desvio-padrão amostral $S(X) = \sqrt{0{,}182} = 0{,}426$.

Questões para reflexão

Você consegue decidir quando deverá utilizar a variância amostral ou a variância populacional em uma pesquisa científica? Pense em alguns exemplos de aplicação.

Atividades de aprendizagem

1. A prefeitura está estudando a possibilidade de implantar um semáforo em um cruzamento da cidade. Para decidir, resolveu considerar o número de acidentes nos últimos 12 meses nesse entroncamento, obtendo os seguintes valores:

5 4 7 8 5 6 4 7 9 7 6 8

Encontre o desvio-padrão populacional.

2. Um banco instalou um caixa eletrônico em um posto de combustível e está observando o número de usuários que vêm utilizando o serviço. Diariamente, o número de clientes que utilizaram o serviço nos últimos 32 dias foi:

15 17 16 15 17 14 17 16 16 17 15
18 14 17 15 14 15 14 15 16 17 18
18 17 15 16 14 18 18 16 15 14.

Calcule o desvio-padrão amostral para essa distribuição.

Para saber mais

Existem muitos livros que tratam da utilização da estatística no dia a dia e da importância em interpretar as informações da melhor maneira possível. Um livro que faz a ponte entre a estatística e o cálculo de risco é Desafio aos Deuses, de Peter L. Bernstein. Fica a dica de leitura!

Seção 3 Correlação e regressão linear

Muitas vezes, ao realizarmos uma pesquisa estatística, não nos interessa apenas o comportamento das variáveis isoladamente, mas, sim, o comportamento de duas ou mais variáveis ao mesmo tempo. Por exemplo:

- A variação no valor do dólar em relação ao real afetou a importação de mercadorias?
- O aumento de investimento em turismo resultou em um aumento no fluxo de turistas?
- A quantidade de água adicionada a determinada marca de cimento afeta a qualidade do concreto resultante?
- O aumento salarial para funcionários de certa empresa afetou o volume de vendas?
- O investimento na compra de maquinário reverteu positivamente no lucro de uma empresa?

Nesta seção, aprenderemos a relacionar duas variáveis e a verificar se há algum tipo de relação entre elas: se o comportamento de uma afeta o comportamento da outra.

3.1 Relação entre variáveis

Suponhamos que uma pesquisa detectou o número de vestibulares prestados por cada estudante antes da sua aprovação em determinada universidade. Para que a pesquisa ficasse mais completa, foi perguntado também a cada um dos estudantes se ele trabalhava na época ou não. Os dados obtidos com a pesquisa aplicada a 10 estudantes foram os seguintes:

X	1	1	2	1	1	2	3	1	1	1
Y	Não	Sim	Não	Não	Não	Sim	Sim	Não	Sim	Sim

onde X é a variável ‘número de vestibulares prestados antes da primeira aprovação’ e Y é a variável ‘trabalhava’. Imagina-se que o fato de trabalhar fora afete o desempenho do estudante nos vestibulares negativamente, ou seja, imagina-se que essas variáveis estejam relacionadas de alguma forma, mas olhando simplesmente para os dados sem agrupá-los fica difícil tirar qualquer conclusão.

Uma tabela de dupla entrada poderia facilitar a análise. Para isso, vamos construir uma tabela auxiliar contendo as possibilidades conjuntas de X e Y e a frequência observada de cada uma:

Tabela 4.11 **Tabela de possibilidades conjuntas**

(X, Y)	**ni**
(1, sim)	3
(1, não)	4
(2, sim)	1
(2, não)	1
(3, sim)	1
(3, não)	0
Total	**10**

Fonte: Dados fictícios.

Agora, estamos em condições de construir uma tabela de dupla entrada:

Tabela 4.12 **Tabela de dupla entrada**

Relação entre vestibulares prestados antes da primeira aprovação e trabalho

Número de vestibulares prestados	**Trabalhava na época**		**Total**
	Sim	**Não**	
1	3	4	7
2	1	1	2
3	1	0	1
Total	5	5	10

A primeira e a última coluna nos dão o que chamamos de tabela marginal de X, e a primeira e a última linha nos dão a tabela marginal de Y.

Tabela 4.13 **Tabelas marginais**

xi	**ni**
1	7
2	2
3	1
Total	**10**

yi	n_i
Sim	5
Não	5
Total	10

Observe que nada mais são do que as distribuições de frequência das variáveis X e Y, respectivamente.

A tabela de dupla entrada permite, então, visualizar o comportamento das duas variáveis ao mesmo tempo. Uma vez que sabemos esboçar a tabela de distribuição conjunta de duas variáveis, a pergunta que se coloca agora é: como saber se o comportamento de uma influencia o comportamento da outra? Será que há relação entre elas? Será que o comportamento de uma depende do comportamento da outra?

Essas perguntas são importantes porque, se as variáveis forem dependentes, podemos explicar como uma delas se comporta em função do desempenho da outra.

Embora difíceis de responder por completo, há algumas maneiras de chegarmos em respostas parciais para essas questões.

3.2 Diagrama de dispersão e a correlação linear

Para iniciarmos nossa discussão, consideremos uma amostra aleatória das notas de Cálculo e Estatística de 12 estudantes do curso de Engenharia Elétrica de determinada universidade:

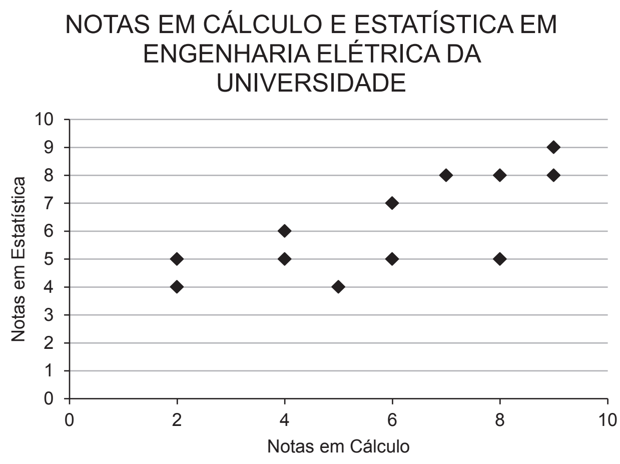
Tabela 4.14 **Gráfico de dispersão**

Estudante	Nota Cálculo	Nota Estatística
1	4	5
2	6	7
3	7	8
4	6	5
5	9	9
6	8	8
7	8	5
8	2	4
9	2	5
10	5	4
11	4	6
12	9	8

Fonte: Novaes e Coutinho (2009).

Queremos saber se há relação entre os desempenhos em Cálculo e Estatística. Para isso, vamos traçar o gráfico de dispersão entre as duas variáveis. O primeiro passo é definirmos qual das variáveis ocupará a posição do eixo das abscissas (eixo horizontal X) e qual ocupará a posição do eixo das ordenadas (eixo vertical Y). Vamos considerar a variável X como sendo 'nota de Cálculo' e a variável Y como sendo 'nota de Estatística'. Assim, os valores associados à nota de Cálculo comporão as coordenadas x e os valores associados à nota de Estatística comporão as coordenadas y no par ordenado (x, y).

Gráfico 4.1 **Gráfico de dispersão**



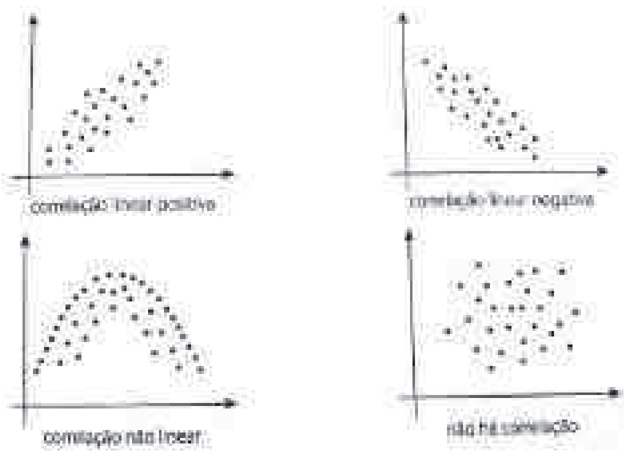
Fonte: Da autora (2014).

O gráfico de dispersão dos dados anteriores nos mostra que há uma relação entre as notas em Estatística e Cálculo: aparentemente, os alunos que possuem maiores notas em Cálculo obtêm melhores notas em Estatística e vice-versa. Assim, podemos arriscar a dizer que as variáveis são correlacionadas.

Segundo Crespo (2009, p.147), podemos dividir os casos de correlação entre variáveis em três:

- Correlação linear positiva: quando o aumento da variável independente X implica em um aumento na variável dependente Y;
- Correlação linear negativa: quando o aumento da variável independente X implica em uma diminuição na variável dependente Y;
- Correlação não linear: quando parece ter algum tipo de relação entre as variáveis em formato de 'curva'.

Figura 4.1 **Tipos de correlação.**



Fonte: Crespo (2005, p. 147).

Mas será que é possível medir essa correlação entre variáveis?

3.3 Coeficiente de correlação

Vamos, agora, apresentar uma maneira de medir a correlação linear entre duas variáveis, chamada de coeficiente de correlação de Pearson, ou coeficiente de correlação linear.

Formalmente, se X e Y são duas variáveis discretas definidas a partir do mesmo fenômeno, com valores atribuídos $(x_i, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, definimos o coeficiente de correlação de Pearson por:

$$r = \frac{n \cdot \left(\sum_{i=1}^{n} x_i \cdot y_i\right) - \left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n \cdot \left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2\right] \cdot \left[n \cdot \left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

Atenção — O símbolo Σ significa somatório, ou seja, $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$

Note que a maior dependência que poderia ocorrer entre duas variáveis X e Y seria X se comportar exatamente como Y: cada aumento da variável X representaria um aumento de mesma quantidade de Y, ou uma diminuição de mesma quantidade de Y. Em termos de correlação, isso significa que $-1 \leq r \leq 1$.

Já no caso de X e Y serem independentes linearmente, r = 0. Segundo Urbano (2010, p. 440), podemos considerar a seguinte escala para correlação:

1,00	Correlação positiva perfeita
0,75	Correlação positiva forte
0,50	Correlação positiva média
0,25	Correlação positiva fraca
0,00	Não existe correlação linear
-0,25	Correlação negativa fraca
-0,50	Correlação negativa média
-0,75	Correlação negativa forte
-1,00	Correlação negativa perfeita

Questões para reflexão

Se o coeficiente de correlação de Person para duas variáveis resultar em 0, podemos concluir que essas variáveis são independentes?

Observe que a fórmula da correlação linear é bastante trabalhosa. Entretanto, podemos utilizar uma tabela auxiliar para ajudar no cálculo, assim como fizemos para calcular a variância. Voltando aos dados da tabela 4.14, e lembrando que X representa a variável 'nota em Cálculo' e Y representa a variável 'nota em Estatística', construímos a seguinte tabela auxiliar:

Tabela 4.15 **Tabela Auxiliar**

i	x_i	y_i	x_i^2	y_i^2	$x_i \cdot n_i$
1	4	5	16	25	20
2	6	7	36	49	42
3	7	8	49	64	56
4	6	5	36	25	30
5	9	9	81	81	81
6	8	8	64	64	64
7	8	5	64	25	40
8	2	4	4	16	8
9	2	5	4	25	10
10	5	4	25	16	20
11	4	6	16	36	24
12	9	8	81	64	72
Total	70	74	476	490	467

Fonte: Da autora (2014).

Da tabela auxiliar, temos:

$$\sum_{i=1}^{12} x_i = 70 \qquad \sum_{i=1}^{12} y_i = 74$$

$$\sum_{i=1}^{12} x_i^2 = 476 \qquad \sum_{i=1}^{12} x_i^2 = 490 \qquad \sum_{i=1}^{12} x_i \cdot\ y_i = 467$$

Agora, podemos calcular o coeficiente:

$$r = \frac{12 \cdot\ 467 - 70 \cdot\ 74}{\sqrt{\left[12 \cdot\ 476 - (70)^2\right] \cdot \left[12 \cdot\ 490 - (74)^2\right]}} = 0,74$$

Observe que o valor encontrado para a correlação é relativamente alto e é positivo. Isso significa que o aumento da variável X 'notas em Cálculo' implica em um aumento considerável da variável Y 'notas em Estatística'.

Vamos observar outro exemplo. A tabela a seguir exibe os dados sobre a quantidade vendida de certa mercadoria e os preços alcançados nos últimos meses.

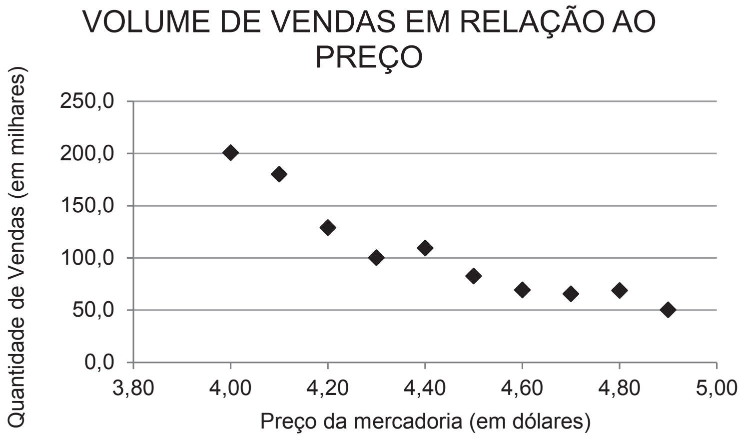
Tabela 4.16 **Número de vendas e preços alcançados**

Mês	Quantidade vendida (em milhares)	Preço (em dólares)
1	50,0	4,9
2	68,4	4,8
3	65,3	4,7
4	69,0	4,6
5	82,3	4,5
6	109,1	4,4
7	99,9	4,3
8	128,6	4,2
9	180,0	4,1
10	200,5	4,0

Fonte: Silver (2000, p. 223).

Vamos montar o gráfico de dispersão para essas variáveis e calcular seu coeficiente de correlação linear. Interessa-nos descrever o comportamento do volume de vendas em relação ao preço — então, o preço fará o papel da variável independente, enquanto o volume de vendas fará o papel da variável dependente.

Gráfico 4.2 Dispersão das vendas em função do preço



Fonte: Da autora (2014).

Aparentemente, há correlação entre as variáveis. E, de fato, se calcularmos a correlação, encontraremos r = - 0,93. Podemos concluir que há uma alta correlação entre as variáveis, sendo que o crescimento de uma (preço) provoca o decrescimento da outra (vendas).

Uma vez que há correlação linear entre as variáveis, nosso objetivo agora será encontrar a equação da reta que melhor descreve essa situação. De posse dessa equação, poderemos utilizá-la pra obter informações que não foram observadas diretamente, mas que podem ser obtidas na análise gráfica. Por exemplo, na pesquisa, não foi mencionada a quantidade de vendas se o preço da mercadoria for cinco dólares, mas se tivermos a equação da reta que melhor aproxima os pontos, poderemos encontrar uma aproximação muito boa para esse valor. Do mesmo modo, se quisermos ter uma ideia da quantidade de vendas esperadas caso o valor da mercadoria for de três dólares e cinquenta centavos, procederemos da mesma maneira.

Vamos, então, aprender as técnicas para encontrarmos a reta que melhor aproxima os pontos observados. Esta técnica é conhecida como regressão linear e o método para encontrá-la recebe o nome de método dos mínimos quadrados.

3.4 Regressão linear

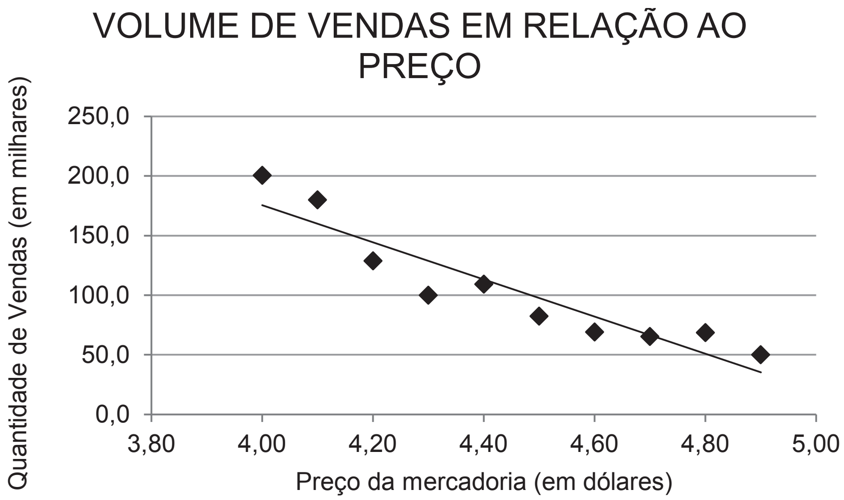
Queremos aproximar os dados do gráfico de dispersão por meio de uma reta cuja equação é da forma $y = \alpha + \beta \cdot x$, onde α é chamado de coeficiente

linear da reta e β é chamado de coeficiente angular da reta. Esses valores α e β são constantes e são eles que a caracterizam: cada reta tem um valor para α e um valor para β fixos. Para traçar a reta, vamos, então, fornecendo valores para x (variável independente) e encontrando um único y tal que $y = \alpha + \beta \cdot x$.

Ícone — É importante lembrar que α e β são duas letras gregas chamadas alfa e beta.

Voltemos, agora, para a nossa situação inicial e vamos traçar uma reta que, aparentemente, aproxima-se de todos os pontos observados da melhor maneira possível (esse procedimento chama-se regressão linear). Provavelmente, nenhum dos pontos (x_i, y_i) da amostra pertence à reta, conforme você pode observar. Assim, para cada x_i da amostra, teremos $y_i = \alpha + \beta \cdot x_i + \varepsilon_i$, onde ε_i é o erro cometido — a distância entre a reta e cada y_i — também chamado de resíduo.

Gráfico 4.3 **Resíduos**



Fonte: Da autora (2014).

O modelo de regressão linear supõe que a média dos valores ε_i deve ser zero. Como consequência,

Ou seja, $\bar{y} = \alpha + \beta \cdot \bar{x}$.

Assim, temos dois valores para serem encontrados, α e β.

Da matemática básica, sabemos que, quando precisamos encontrar duas incógnitas, precisamos de um sistema com duas equações que as envolvam. O modelo de regressão linear nos forneceu uma das equações: $\bar{y} = \alpha + \beta \cdot \bar{x}$. Precisamos de mais uma relação entre α e β para podermos encontrá-los.

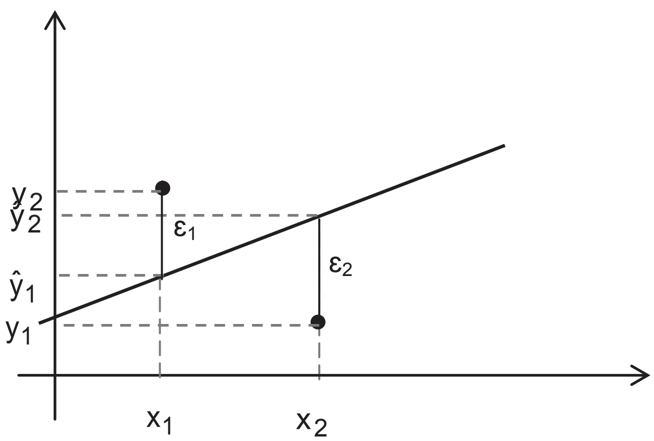
Para que a média dos valores ε_i seja zero, temos que $\frac{(\varepsilon_1 + \varepsilon_2 + ... + \varepsilon_n)}{n} = 0$. Como n é o tamanho da amostra e, portanto, não é zero, $(\varepsilon_1 + \varepsilon_2 + ... + \varepsilon_n) = 0$. Mas, para cada i, quem é ε_i?

Uma vez que $y = \alpha + \beta \cdot x_1 + \varepsilon_1$, se considerarmos $\hat{y} = \alpha + \beta \cdot x_1$, veremos que $\varepsilon_1 + y_i - \hat{y}_i$ para cada i. Portanto, se a soma dos resíduos deve ser zero,

$$(y_i - \hat{y}_i) + (y_2 - \hat{y}_2) + ... + (y_n - \hat{y}_n) = 0$$

Mas esse fato não garante que a reta seja a mais próxima possível, pois o resíduo encontrado para certo i pode anular um resíduo encontrado para certo j (i ≠ j). Observe a situação:

Gráfico 4.4 Função



Fonte: Da autora (2014).

Observe o desenho: os dois resíduos, embora grandes, possuem tamanho similar, mas sinais diferentes. Logo, quando forem somados, irão resultar em um número próximo de zero.

Para contornar esse problema, em vez de minimizar a soma dos resíduos, vamos minimizar a soma do quadrado dos resíduos. Assim, teremos apenas valores positivos envolvidos no cálculo e, para que a soma resulte em um valor próximo de zero, cada um dos resíduos precisa ser muito pequeno.

$\left(\varepsilon_1^2 + \varepsilon_2^2 + ... + \varepsilon_n^2\right) = 0$, ou seja, $(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + ... + (y_n - \hat{y}_n)^2 = 0$

O critério acima é conhecido como método dos mínimos quadrados.

Através dessa suposição e através de cálculos matemáticos, chega-se nas seguintes equações:

$$\beta = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}$$

$$\alpha = \overline{y} - \hat{\beta} \cdot \overline{x}$$

Ou, de maneira mais condensada:

$$\beta = \frac{n \cdot \sum X \cdot Y - \sum X \cdot \sum Y}{n \cdot \sum X^2 - \left(\sum X \right)^2}$$

$$\alpha = \overline{Y} - \hat{\beta} \cdot \overline{X}$$

Atenção — Como estamos utilizando dados de uma amostra para encontrar a equação da reta que melhor aproxime os dados, na verdade, estamos estimando a verdadeira equação da reta. Por essa razão, escrevemos $\hat{Y} = \alpha + \beta \cdot X$.

Vamos, então, calcular a regressão linear para o problema do volume de vendas em relação ao preço, utilizando o método dos mínimos quadrados. A mesma tabela auxiliar que montamos para calcular a correlação nos ajuda no cálculo das variáveis α e β.

Tabela 4.17 **Tabela Auxiliar**

i	x_i (preço)	y_i (vendas)	x_i^2	y_i^2	$x_i \cdot n_i$
1	4,9	50,0	24,0	2500,0	245,0
2	4,8	68,4	23,0	4678,6	328,3
3	4,7	65,3	22,1	4264,1	306,9
4	4,6	69,0	21,2	4761,0	317,4
5	4,5	82,3	20,3	6773,3	370,4
6	4,4	109,1	19,4	11902,8	480,0
7	4,3	99,9	18,5	9980,0	429,6
8	4,2	128,6	17,6	16538,0	540,1
9	4,1	180,0	16,8	32400,0	738,0
10	4,0	200,5	16,0	40200,3	802,0
Total	44,5	1053,1	198,9	133998,0	4557,7

Fonte: Da autora (2014).

Assim,

$$\beta = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{10 \cdot 4557,7 - 44,5 \cdot 1053,1}{10 \cdot 198,9 - (44,5)^2} = -155,9$$

$$\alpha = \bar{y} - \beta \cdot \bar{x} = \frac{1053,1}{10} - (-155,9) \cdot \frac{44,5}{10} = 798,9$$

Assim, a reta que melhor aproxima os dados da amostra é dada pela equação $\hat{Y} = 798,9 - 155,9 \cdot X$.

Exemplo: Consideremos a situação proposta no tópico anterior, em que eram analisadas as notas em Estatística e em Cálculo obtidas pelos estudantes de Engenharia Elétrica de uma determinada universidade. Vimos que as variáveis X – nota em Cálculo — e Y — nota em Estatística — estavam correlacionadas linearmente. Vamos, então, utilizar o método de minimização dos quadrados para encontrar a reta que melhor aproxima os dados da amostra. Relembrando a tabela auxiliar 4.15 que construímos para calcular a correlação entre as variáveis era a seguinte:

Tabela 4.15 Tabela Auxiliar

i	x_i	y_i	x_i^2	y_i^2	$x_i \cdot n_i$
1	4	5	16	25	20
2	6	7	36	49	42
3	7	8	49	64	56
4	6	5	36	25	30
5	9	9	81	81	81
6	8	8	64	64	64
7	8	5	64	25	40
8	2	4	4	16	8
9	2	5	4	25	10
10	5	4	25	16	20
11	4	6	16	36	24
12	9	8	81	64	72
Total	70	74	476	490	467

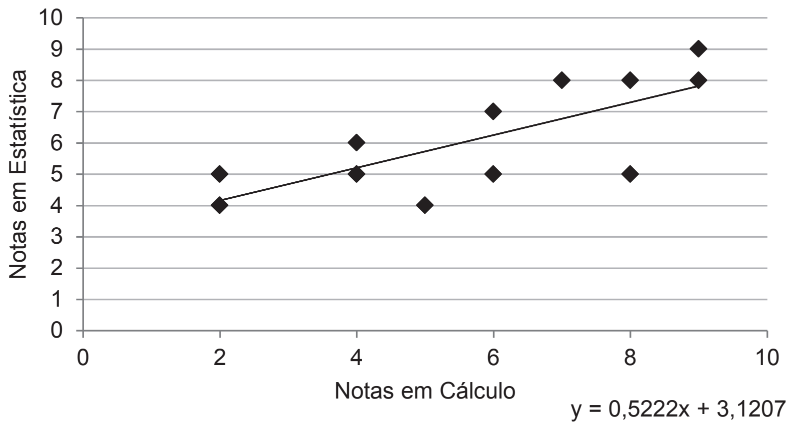
Fonte: Da autora (2014).

$$\beta = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} = \frac{12 \cdot 467 - 70 \cdot 74}{12 \cdot 476 - (70)^2} = 0{,}522$$

$$\alpha = \overline{y} - \beta \cdot \overline{x} = \frac{74}{12} - (0{,}522) \cdot \frac{70}{12} = 3{,}121$$

Assim, a reta que melhor aproxima os dados da amostra é dada pela equação $\hat{Y} = 3{,}121 + 0{,}522 \cdot X$.

Gráfico 4.5 Notas dos estudantes de engenharia elétrica da universidade



Fonte: Da autora (2014).

A regressão linear é, portanto, a reta que melhor aproxima os dados de uma amostra onde duas variáveis são analisadas conjuntamente. O que isso significa?

Voltemos ao exemplo anterior das notas em Cálculo e em Estatística. O modelo de regressão linear nos forneceu a seguinte equação: $\hat{Y} = 3{,}121 + 0{,}522 \cdot X$. Note que o valor de α — no caso, α = 3,121 — independe de X. Isso significa que, mesmo sem nota em Cálculo, é de se esperar que o aluno obtenha nota 3,122, ou seja, é a nota inicial em Estatística segundo o modelo. Já a constante β — no caso, β = 0,522 — nos fornece a proporção com que Y varia quando X varia. Nesse caso, cada alteração na nota de Cálculo (X) altera a nota de Estatística a uma proporção de 0,522 (ou 52,20%).

É claro que a reta não descreve exatamente o comportamento da relação entre as variáveis, mas dá uma estimativa para o comportamento conjunto delas. No exemplo anterior, das notas em Cálculo e Estatística, se procurarmos a nota de Estatística (variável Y) correspondente à nota 6 em Cálculo (variável X) na reta, encontraremos $\hat{Y} = 3{,}121 + 0{,}522 \cdot 6 = 6{,}25$, ou seja, de acordo com a equação da reta, quem tem nota 6 em Cálculo obtém nota 6,25 em Estatística. Por outro lado, na amostra, dois estudantes tiraram 6 em Cálculo: um tirou 5 em Estatística e o outro tirou 7! O fato dos valores observados serem diferentes do obtido via regressão não significa que houve erro no cálculo, pois são informações diferentes: o dado obtido via equação é uma estimativa.

Esse exemplo ilustra o cuidado que devemos ter ao trabalharmos com a linearização: a regressão é uma estimativa que descreve o comportamento conjunto dos dados, mas não significa que podemos replicar os dados da amostra com ela.

Por outro lado, a equação é bastante útil para tirarmos informações sobre dados dos quais não dispomos.

Suponhamos que estivéssemos interessados em obter uma estimativa para a nota de um estudante que obteve 4,5 em Cálculo. Nenhum dos estudantes que fez parte da amostra obteve essa nota e, portanto, não podemos utilizar informações da tabela de distribuição. Por outro lado, podemos estimar um valor para ela através da equação da reta:

$$\hat{Y} = 3{,}121 + 0{,}522 \cdot X = 3{,}121 + 0{,}522 \cdot 4{,}5 = 5{,}47$$

Assim, é estimado que esse estudante tenha tirado 5,47 ou 5,5 em Estatística.

Note que, na amostra, os dados de x variavam de 2 a 9, ou seja, x pertence ao intervalo fechado [2, 9]. Portanto, embora não faça parte da amostra, x = 4,5 também pertence a esse intervalo, pois é maior do que 2 e menor do que 9. Quando o valor y que pretendemos estimar estiver associado a um valor x pertencente ao intervalo de valores da amostra, damos a esse processo o nome de **interpolação**.

E se quiséssemos estimar a nota em Estatística de um estudante que obteve 10 em Cálculo? Embora o valor 10 não faça parte do intervalo de valores para x da amostra, o fato de estarmos aproximando os dados por uma reta (regressão linear) implica em podermos estimar a nota em Estatística via equação. Nesse caso, $\hat{Y} = 3{,}121 + 0{,}522 \cdot X = 3{,}121 + 0{,}522 \cdot 10 = 8{,}34$.

Assim, é estimado que o estudante que tenha nota 10 em Cálculo tenha nota 8,3 em Estatística.

Quando o valor y que pretendemos estimar estiver associado a um valor x não pertencente ao intervalo de valores da amostra, damos a esse processo o nome de **extrapolação**.

Para a interpolação, qualquer valor estudado é aceitável, uma vez que ele pertence ao intervalo da amostra; já no caso da extrapolação, precisamos ter alguns cuidados. Embora a reta esteja definida para qualquer número real (propriedade de reta), o significado da variável impõe restrições aos valores estudados.

Questões para reflexão

Você consegue pensar em alguma restrição para a extrapolação das notas dos estudantes?

Exemplo: Vamos voltar ao exemplo do volume de vendas associado ao preço da mercadoria.

Tabela 4.16 **Número de vendas e preços alcançados**

Mês	Quantidade vendida (em milhares)	Preço (em dólares)
1	50,0	4,9
2	68,4	4,8
3	65,3	4,7
4	69,0	4,6
5	82,3	4,5
6	109,1	4,4
7	99,9	4,3
8	128,6	4,2
9	180,0	4,1
10	200,5	4,0

Fonte: Silver (2000, p. 223).

Já realizamos a regressão linear para essa situação e encontramos a seguinte equação da reta: $\hat{Y} = 798{,}9 - 155{,}9 \cdot X$.

Se quisermos estimar o volume de vendas caso o preço da mercadoria seja de 6 dólares, teremos que

$$\hat{Y} = 798{,}9 - 155{,}9 \cdot X = 798{,}9 - 155{,}9 \cdot 6 = -136{,}27 .$$

Note que, embora os cálculos estejam corretos, o valor encontrado é negativo, o que não faz sentido, uma vez que a variável Y está associada a quantidades. Por outro lado, fica evidente que cobrar 6 dólares pela mercadoria seria inviável. Vamos encontrar o valor em dólares para o qual a quantidade vendida seria nula?

Nesse caso, y = 0.

$$\hat{Y} = 798{,}9 - 155{,}9 \cdot X \xrightarrow{\hat{Y}=0} 0 = 798{,}9 - 155{,}9 \cdot X$$

$$\begin{aligned} 0 &= 798{,}9 - 155{,}9 \cdot X \\ 155{,}9 \cdot X &= 798{,}9 \\ X &= 5{,}12 \end{aligned}$$

Assim, concluímos que a mercadoria não pode custar mais do que 5,12 dólares.

Atividades de aprendizagem

1. O custo mensal de manutenção de determinado equipamento em uma indústria está sendo analisado em função da sua idade. Nove equipamentos adquiridos em diferentes anos tiveram o custo averiguado, e os dados obtidos foram os seguintes:

Tabela 4.18 **Idade do equipamento em relação ao custo mensal**

Idade do equipamento (anos)	Custo mensal (reais)
1	8
2	13
3	18
4	28
5	24
6	26
7	29
8	32
9	37

a) Trace o gráfico de dispersão.
b) Faça a regressão linear e encontre a equação da reta melhor ajustada.
c) Com base no modelo de regressão linear, qual o custo mensal de um equipamento com 12 anos de uso?

2. Uma empresa submete seus novos operários a um teste de aptidão (X) e três meses depois mede a produtividade desses funcionários (Y). Os resultados estão na tabela a seguir:

Tabela 4.19 **Relação entre a aptidão e produtividade dos funcionários da empresa**

Funcionários	Aptidão(X)	Produtividade (Y)
A	22	45
B	25	37
C	15	25
D	19	40
E	22	33
F	18	30

a) Faça o diagrama de dispersão e calcule o coeficiente de correlação.
b) Encontre a equação da reta de regressão.
c) Para um indivíduo cujo resultado no teste de aptidão foi 20, qual foi a produtividade esperada?
d) Para um indivíduo que obteve 28 no teste de produtividade, qual foi o resultado no teste de aptidão?

Para saber mais

Algumas vezes, os problemas que queremos estudar envolvem mais de duas variáveis, e precisamos saber se a variação de uma influencia linearmente na variação das outras. Nesses casos, a regressão múltipla é utilizada, consistindo em encontrar uma equação matemática (no caso, a equação de um plano) que descreva o comportamento de uma variável em função de várias outras. Assim, a regressão linear é como um caso particular da regressão múltipla. Caso você tenha interesse em seguir seus estudos, consulte o livro de Urbano, que faz parte das referências bibliográficas deste material.

Fique ligado!

Nesta unidade, você aprendeu que:

- A média aritmética e a média ponderada fornecem uma medida de posição que resume os dados obtidos para determinada variável.
- A variância e o desvio-padrão nos informam como os dados obtidos estão distribuídos, ou dispersos, em torno da média obtida.
- Existem a variância amostral e a variância populacional. Enquanto uma deve ser calculada nos estudos envolvendo amostras, a outra é utilizada quando trabalhamos com toda a população.
- O coeficiente de correlação linear nos dá uma medida da dependência linear entre duas variáveis, e seu número deve estar entre -1 e 1.
- A regressão linear é a reta que melhor aproxima os dados observados para duas variáveis em um gráfico de dispersão entre elas.
- Os valores α e β são constantes; são eles que caracterizam a reta e, portanto, são eles que devem ser encontrados.
- O método dos mínimos quadrados consiste em um método matemático que permite encontrar α e β, de modo que a soma dos quadrados dos resíduos seja a menor possível.
- Interpolação é o processo de, dado um valor para x pertencente ao intervalo da amostra, encontrar um valor estimado para y via regressão linear.
- Extrapolação é o processo de, dado um valor para x não pertencente ao intervalo da amostra, encontrar um valor estimado para y via regressão linear.
- Sempre é possível interpolar, mas é preciso cuidado na hora de extrapolar!

Para concluir o estudo da unidade

Caro estudante,

A unidade que você acabou de estudar é uma introdução aos conteúdos apresentados. Existem outras medidas resumo, assim como é possível ampliar a ideia da regressão linear entre duas variáveis para mais de duas variáveis através da regressão múltipla. Você pode aprofundar seu estudo consultando qualquer um dos livros didáticos que constam na bibliografia. Faça isso! Não se contente apenas com o que lhe é apresentado, mas prossiga seus estudos.

Atividades de aprendizagem da unidade

1. Em uma clínica cardíaca, foram anotados os níveis de colesterol (em mg/100ml) para trinta pacientes, homens com idade entre 40 e 60 anos que foram à clínica fazer um *check-up*.

Paciente	1	2	3	4	5	6	7	8	9	10
Colesterol	160	160	161	163	167	170	172	172	173	177

Paciente	11	12	13	14	15	16	17	18	19	20
Colesterol	178	181	181	182	185	186	194	197	199	203

Paciente	21	22	23	24	25	26	27	28	29	30
Colesterol	203	205	206	206	208	209	211	214	218	225

Calcule o nível de colesterol médio e o desvio-padrão amostral.

2. Um posto de saúde de certo bairro mantém um arquivo com o número de pacientes que procuram o consultório odontológico diariamente. Os dados obtidos no último mês foram os seguintes:

3	4	3	4	5	1	6	3	4	5	3	4	3
3	4	3	5	5	5	5	6	11	10	2	1	2
3	1	5	2.									

Calcule a variância populacional para essa distribuição.

3. Adaptado de Novaes e Coutinho (2009). Uma agência de turismo especializada em oferecer passeios opcionais para turistas que visitam determinada região está estudando a variação na adesão a determinado pacote quando são acrescentados ou tirados percursos do preço cobrado. Eles obtiveram as seguintes informações:

Tabela 4.19 **Número de adesões ao pacote em relação ao preço**

Preço ($)	Numero de adesões
10	50
15	51
20	48
25	43
30	42
35	45
40	39
45	38
50	40
55	34
60	32
70	30
90	25

a) Faça o gráfico de dispersão.

b) Há correlação linear entre os dados? Justifique.

c) Encontre a função matemática que explique a dependência entre o número de adesões e o preço do passeio opcional.

d) Estime o número de pessoas que farão o passeio opcional se o valor cobrado for de R$ 80,00.

4. Adaptado de Crespo (2005). Certa empresa, estudando a variação da demanda do seu produto em relação à variação de preço de venda (em unidades monetárias), obteve os seguintes dados:

Tabela 4.20 **Demanda do produto em relação ao preço de venda**

Preço (u.m.)	Demanda
38	350
42	325
50	297
56	270
59	256
63	246
70	238
80	223
95	215
110	208

a) Determine os coeficientes de correlação entre as variáveis.
b) Encontre a equação da reta ajustada.
c) Se o preço de venda for 75 u.m., qual será a demanda estimada?
d) Se o preço de venda for de 110 u.m., qual será a demanda estimada?
e) Qual o valor máximo de preço possível para que haja demanda?

5. Adaptado de Magalhães (2010). Uma indústria submete seus novos operários a um teste de aptidão (X) e três meses depois mede a produtividade desses operários (Y). Os resultados estão na tabela a seguir:

Tabela 4.20 **Relação entre Aptidão e Produtividade dos Funcionários**

Operário	Aptidão(X)	Produtividade (Y)
A	22	45
B	25	37
C	15	25
D	19	40
E	22	33
F	18	30

a) Encontre a equação da reta de regressão.
b) Para um indivíduo cujo resultado no teste de aptidão foi 20, qual é a produtividade esperada?
c) Para um indivíduo que obteve 28 no teste de produtividade, qual foi o resultado no teste de aptidão?

Referências

CRESPO, A. A. **Estatística fácil**. 19. ed. São Paulo: Saraiva, 2005.

MAGALHÃES, M. N.; LIMA, A. C. P. **Noções de probabilidade e estatística**. São Paulo: Edusp, 2010.

NOVAES, D. V; COUTINHO, S.; QUEIROZ, C. **Estatística para educação profissional**. São Paulo: Atlas, 2009.

SILVER, M. **Estatística para administração**. São Paulo: Atlas, 2000.

URBANO, J. O. **Estatística**: uma nova abordagem. Rio de Janeiro: Ciência Moderna, 2010.